AWS Well-Architected Framework

# Pilar Eficiência de performance

# Pilar Eficiência de performance: AWS Well-Architected Framework

# Table of Contents

# Pilar Eficiência de performance: AWS Well-Architected Framework

Data de publicação: 3 de outubro de 2023 (Revisões do documento)

## Resumo

Este whitepaper enfoca o pilar Eficiência de performance do AWS Well-Architected Framework. O escopo deste documento é fornecer orientações que ajudem os clientes a usar os recursos da nuvem de forma eficiente para atender às necessidades de seus negócios e manter essa eficiência à medida que a demanda muda e as tecnologias evoluem.

## Introdução

O AWS Well-Architected Framework ajuda a compreender os prós e os contras das decisões tomadas ao criar workloads na AWS. O uso do Framework ajuda você a aprender as práticas de arquitetura recomendadas para projetar e operar workloads confiáveis, seguras, eficientes, econômicas e sustentáveis na nuvem. Ele fornece uma maneira de você avaliar consistentemente suas arquiteturas em relação às melhores práticas e identificar áreas de aprimoramento. Acreditamos que ter as cargas de trabalho bem arquitetadas aumenta muito a probabilidade de sucesso nos negócios.

A estrutura é baseada em seis pilares:

- Excelência Operacional
- Segurança
- Confiabilidade
- Eficiência de performance
- Otimização de custos
- Sustentabilidade

Este documento enfoca a aplicação dos princípios do pilar de eficiência de performance às suas workloads. Em ambientes tradicionais on-premises, alcançar uma performance elevada e duradoura é algo desafiador. O uso dos princípios apresentados neste documento ajudará você a criar

arquiteturas na AWS que entreguem, com eficácia, performance constante ao longo do tempo. A orientação e as práticas recomendadas deste documento estão distribuídas em cinco áreas de foco principais que servem como princípios orientadores para a criação de soluções de nuvem eficientes na AWS. Essas áreas de foco são:

- Seleção de arquitetura
- Computação e hardware
- Gerenciamento de dados
- Rede e entrega de conteúdo
- Processo e cultura

Este documento é destinado a pessoas que ocupam cargos de tecnologia, como diretores de tecnologia (CTOs), arquitetos, desenvolvedores e membros da equipe de operações. Depois de ler este documento, você entenderá as práticas recomendadas e as estratégias da AWS a serem usadas ao projetar arquiteturas de nuvem de alta performance.

# Eficiência de performance

O pilar Eficiência de performance tem como foco o uso eficiente de recursos de computação para atender a requisitos e os meios para manter essa eficiência conforme a demanda muda e as tecnologias evoluem.

Tópicos



## Princípios de design

Os princípios de design a seguir podem ajudar você a alcançar e manter workloads eficientes na nuvem.

- Democratize tecnologias avançadas: facilite a implementação de tecnologia avançada para a sua equipe delegando tarefas complexas ao seu fornecedor de nuvem. Em vez de solicitar que sua equipe de TI aprenda sobre como hospedar e executar uma nova tecnologia, avalie a possibilidade de consumir a tecnologia como um serviço. Por exemplo, bancos de dados NoSQL, transcodificação de mídia e machine learning são tecnologias que exigem altos níveis de especialização. Na nuvem, essas tecnologias se tornam serviços que sua equipe pode consumir, permitindo que a equipe se concentre no desenvolvimento de produtos, em vez de provisionamento e gerenciamento de recursos.
- Torne-se global em poucos minutos: a implantação da workload em várias Regiões da AWS em todo o mundo permite que você forneça uma latência menor e uma experiência melhor para os clientes a um custo mínimo.
- Use arquiteturas sem servidor: as arquiteturas sem servidor eliminam a necessidade de executar e manter servidores físicos para realizar atividades tradicionais de computação. Os serviços de armazenamento sem servidor, por exemplo, podem atuar como sites estáticos (eliminando a necessidade de servidores da web) e os serviços de eventos podem hospedar o código. Isso elimina o fardo operacional do gerenciamento de servidores físicos e pode reduzir os custos transacionais, pois os serviços gerenciados operam em escala de nuvem.
- Experimentar com mais frequência: Com recursos virtuais e automatizáveis, você pode executar rapidamente testes comparativos usando diferentes tipos de instâncias, armazenamento ou configurações.

- Considerar a afinidade mecânica: use a abordagem tecnológica que se alinhe melhor às suas metas. Por exemplo, avalie padrões de acesso a dados ao selecionar banco de dados ou armazenamento para a workload.

# Definição

Concentre-se nas seguintes áreas para alcançar eficiência de performance na nuvem:

- Seleção de arquitetura
- Computação e hardware
- Gerenciamento de dados
- Rede e entrega de conteúdo
- Processo e cultura

Adote uma abordagem baseada em dados para criar uma arquitetura de alto desempenho. Reúna dados sobre todos os aspectos da arquitetura, desde o design de alto nível até a seleção e a configuração dos tipos de recursos.

Analisar suas escolhas regularmente garante que você está tirando proveito da evolução contínua da Nuvem AWS. O monitoramento garante que você esteja ciente de qualquer desvio em relação à performance esperada. Faça concessões em sua arquitetura visando o aprimoramento da performance, como o uso de compactação ou armazenamento em cache, ou ainda a diminuição dos requisitos de consistência.

# Seleção de arquitetura

A solução ideal para uma workload específica varia e, muitas vezes, as soluções combinam várias abordagens. Workloads do Well-Architected usam várias soluções e permitem diferentes recursos para aprimorar a performance.

Os recursos da AWS estão disponíveis em vários tipos e configurações, facilitando encontrar uma abordagem que atenda melhor às suas necessidades. Você também pode encontrar opções que não são facilmente obtidas com infraestrutura on-premises. Por exemplo, um serviço gerenciado. como o Amazon DynamoDB, fornece um banco de dados NoSQL totalmente gerenciado com latência de milissegundos de um dígito em qualquer escala.

Essa área de foco compartilha orientações e práticas recomendadas sobre como selecionar padrões de arquitetura e recursos de nuvem eficientes e de alto desempenho.

Práticas recomendadas

- PERF01-BP01 Conheça e compreenda os serviços e recursos de nuvem disponíveis
- PERF01-BP02 Use a orientação de seu provedor de nuvem ou de um parceiro apropriado para aprender sobre padrões de arquitetura e práticas recomendadas
- PERF01-BP03 Inclua o custo nas decisões de arquitetura
- PERF01-BP04 Avalie como certas trocas (trade-offs) afetam os clientes e a eficiência da arquitetura
- PERF01-BP05 Use políticas e arquiteturas de referência
- PERF01-BP06 Use testes comparativos para orientar decisões de arquitetura
- PERF01-BP07 Use uma abordagem baseada em dados para escolhas de arquitetura

## PERF01-BP01 Conheça e compreenda os serviços e recursos de nuvem disponíveis

Continue a descobrir e aprender sobre serviços e configurações disponíveis que ajudam a tomar decisões e melhorar a eficiência da performance de suas workloads com base na arquietura.

Antipadrões comuns:

- Você usa a nuvem como um datacenter colocalizado.

- Você não moderniza sua aplicação após a migração para a nuvem.
- Você só usa um tipo de armazenamento para tudo que precisa ser mantido.
- Você usa tipos de instância mais próximos aos padrões atuais, no entanto, maiores, quando necessário.
- Você implanta e gerencia tecnologias disponíveis como serviços gerenciados.

Benefícios de estabelecer esta prática recomendada: Ao pensar em novos serviços e configurações, você poderá melhorar consideravelmente a performance, reduzir custos e otimizar o esforço necessário para manter as workloads. Isso também pode ajudar a acelerar o tempo para valorização dos produtos habilitados para a nuvem.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

A AWS lança constantemente novos serviços e recursos que podem melhorar a performance e reduzir o custo das workloads na nuvem. Atualizar-se com relação a esses novos serviços e atributos é crucial para manter a eficácia da performance na nuvem. Modernizar a arquitetura da workload também ajuda a acelerar a produtividade, impulsionar a inovação e ter acesso a mais oportunidades de crescimento.

### Etapas da implementação

- Faça um inventário do software e da arquitetura usados para serviços relacionados a suas workloads. Decida sobre qual categoria de produtos você quer saber mais.
- Explore as ofertas da AWS para identificar e aprender sobre os serviços e as opções de configuração relevantes que podem ajudar você a melhorar a performance e reduzir os custos e a complexidade operacional.
  - Quais são as novidades da AWS?
  - Blog da AWS
  - AWS Skill Builder
  - Eventos e webinars da AWS
  - Treinamento da AWS and Certifications
  - Canal da AWS no Youtube
  - Workshops da AWS

  - Comunidades da AWS
- Use ambientes sandbox (sem produção) para aprender e experimentar novos serviços sem incorrer em custos adicionais.

## Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS
- Crie aplicações modernas na AWS

Vídeos relacionados:

- This is my Architecture

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

# PERF01-BP02 Use a orientação de seu provedor de nuvem ou de um parceiro apropriado para aprender sobre padrões de arquitetura e práticas recomendadas

Use recursos disponibilizados pelo fornecedor de nuvem, como documentação, arquitetos de soluções, serviços profissionais ou parceiros apropriados, para orientar suas decisões durante a escolha da arquitetura. Eles ajudarão a analisar e melhorar sua arquitetura para alcançar a performance ideal.

Antipadrões comuns:

- Você usa a AWS como um provedor de nuvem comum.

- Você usa as ofertas da AWS de uma maneira para a qual elas não foram projetadas.
- Você segue todas as orientações sem considerar seu contexto de negócios.

Benefícios de estabelecer esta prática recomendada: Usar a orientação de um provedor de nuvem ou de um parceiro apropriado pode ajudar a fazer as escolhas de arquitetura certas para as workloads e ter confiança em suas decisões.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

A AWS oferece uma ampla variedade de orientações, documentações e recursos que podem ajudar a criar e gerenciar workloads eficientes na nuvem. A documentação da AWS fornece exemplos de código, tutoriais e explicações detalhadas do serviço. Além da documentação, a AWS fornece programas de treinamento e certificação, arquitetos de soluções e serviços profissionais que podem ajudar os clientes a explorar diferentes aspectos dos serviços em nuvem e implementar uma arquitetura de nuvem eficiente na AWS.

Aproveite esses recursos para obter informações sobre conhecimentos valiosos e práticas recomendadas, economizar tempo e obter melhores resultados na Nuvem AWS.

### Etapas da implementação

- Analise a documentação e as orientações da AWS e siga as práticas recomendadas. Esses recursos podem ajudar a escolher e configurar serviços com eficiência e obter melhor performance.
  - Documentação da AWS (como guias do usuário e whitepapers)
  - Blog da AWS
  - Treinamento da AWS and Certifications
  - Canal da AWS no Youtube
- Participe de eventos de parceiros da AWS (como Conferências Globais da AWS, AWS re:Invent, grupos de usuários e workshops) para ouvir dos próprios especialistas da AWS quais são as práticas recomendadas no uso dos serviços da empresa.
  - Eventos e webinars da AWS
  - Workshops da AWS
  - Comunidades da AWS

- Entre em contato com a AWS para obter assistência quando precisar de mais orientações ou informações sobre produtos. Os arquitetos de soluções da AWS e o AWS Professional Services fornecem orientação para a implementação da solução. parceiros da AWS oferecem toda a experiência na AWS para ajudar você a adquirir agilidade e inovação para os negócios.
- Use AWS Support se precisar de suporte técnico para otimizar o uso de um serviço. Nossos planos de suporte são projetados a fim de oferecer a combinação certa de ferramentas e acesso ao conhecimento especializado para ter sucesso com a AWS e melhorar a performance, gerenciar riscos e manter os custos sob controle.

## Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS
- AWS Enterprise Support

Vídeos relacionados:

- This is my Architecture

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

# PERF01-BP03 Inclua o custo nas decisões de arquitetura

Considere o custo em suas decisões de arquitetura para melhorar a utilização de recursos e a eficiência da performance de suas workloads na nuvem. Quando você está ciente das implicações de custo de suas workloads na nuvem, é mais provável que utilize recursos eficientes e reduza práticas ineficazes.

Antipadrões comuns:

- Você só usa uma família de instâncias.
- Você não avalia soluções licenciadas em relação a soluções de código aberto.
- Você não define políticas de ciclo de vida de armazenamento.
- Você não analisa os novos serviços e recursos da Nuvem AWS.
- Você só usa o armazenamento em bloco.

Benefícios de estabelecer esta prática recomendada: Levar em conta o custo em sua tomada de decisão permite que você use recursos mais eficientes e examine outros investimentos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

Otimizar as workloads em função do custo pode melhorar a utilização dos recursos e evitar o desperdício em uma workload na nuvem. A consideração do custo nas decisões de arquitetura geralmente inclui o dimensionamento correto dos componentes da workload e a viabilização da elasticidade, o que resulta em maior eficiência da sua performance na nuvem.

### Etapas da implementação

- Estabeleça objetivos de custo, como limites orçamentários para a workload na nuvem.

- Identifique os principais componentes (como instâncias e armazenamento) que impulsionam o custo da workload. Você pode usar o AWS Pricing Calculator e o AWS Cost Explorer para identificar os principais fatores de custo na workload.

- Use práticas recomendadas de otimização de custos do Well-Architected para otimizar esses componentes principais em termos de custo.
- Monitore e analise constantemente os custos para identificar oportunidades de otimizar as workloads e economizar.
  - Use o AWS Budgets para receber alertas quando os custos forem inaceitáveis.
  - Use AWS Compute Optimizer ou AWS Trusted Advisor para receber recomendações de otimização de custos.
  - Use Detecção de Anomalias de Custos da AWS para obter detecção automática de anomalias de custo e análise da causa raiz.

## Recursos

Documentos relacionados:

- A Detailed Overview of the Cost Intelligence Dashboard
- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS

Vídeos relacionados:

- This is my Architecture
- Optimize performance and cost for your AWS compute

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS
- Rightsizing with Compute Optimizer and Memory utilization enabled
- AWS Compute Optimizer Demo code

# PERF01-BP04 Avalie como certas trocas (trade-offs) afetam os clientes e a eficiência da arquitetura

Ao avaliar melhorias relacionadas ao desempenho, determine quais escolhas afetam os clientes e a eficiência das workloads. Por exemplo, se o uso de um datastore de chave-valor aumentar o desempenho do sistema, é importante avaliar como a mudança afetará os clientes após tornar-se consistente.

Antipadrões comuns:

- Você pressupõe que todos os ganhos de desempenho devem ser implementados, mesmo que seja preciso fazer certas trocas para implementação.

- Você só avalia alterações nas workloads quando um problema de performance atinge um ponto crítico.

Benefícios de estabelecer esta prática recomendada: Ao avaliar possíveis melhorias relacionadas à performance, você deve decidir se as concessões para as alterações são aceitáveis com os requisitos da workload. Em alguns casos, pode ser necessário implementar controles adicionais para compensar as compensações.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

Identifique áreas críticas na arquitetura em termos de desempenho e impacto para o cliente. Determine como você pode promover aprimoramentos, quais concessões esses aprimoramentos exigem e como elas afetam o sistema e a experiência do usuário. Por exemplo, a implementação de armazenamento de dados em cache pode ajudar a aprimorar drasticamente a performance, mas requer uma estratégia clara de como e quando atualizar ou invalidar dados em cache a fim de prevenir comportamentos incorretos do sistema.

### Etapas da implementação

- Entenda SLAs e requisitos de suas workloads.
- Defina claramente os fatores de avaliação. Os fatores podem estar relacionados a custo, confiabilidade, segurança e desempenho de suas workloads.
- Selecione arquitetura e serviços que possam atender às suas necessidades.
- Realize experiências e provas de conceitos (POCs) para avaliar os fatores e o impacto de certas trocas para os clientes e para a eficiência da arquitetura. Normalmente, workloads de alta disponibilidade, com bom desempenho e seguras consomem mais recursos da nuvem e, ao mesmo tempo, proporcionam uma melhor experiência ao cliente.

## Recursos

Documentos relacionados:

- Amazon Builders' Library
- Amazon QuickSight KPIs
- Amazon CloudWatch RUM

- Documentação do X-Ray
- Understand resiliency patterns and trade-offs to architect efficiently in the cloud

Vídeos relacionados:

- Build a monitoring plan
- Optimize applications through Amazon CloudWatch RUM
- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Measure page load time with Amazon CloudWatch Synthetics (Medição do tempo de carga da página com o Amazon CloudWatch Synthetics)
- Amazon CloudWatch RUM Web Client (Cliente da web do Amazon CloudWatch RUM)

# PERF01-BP05 Use políticas e arquiteturas de referência

Use políticas internas e arquiteturas de referência existentes ao selecionar serviços e configurações para ser mais eficiente ao projetar e implementar a workload.

Antipadrões comuns:

- Você permite uma ampla variedade de tecnologias que podem afetar os custos de gerenciamento da empresa.

Benefícios de estabelecer esta prática recomendada: Estabelecer uma política para opções de arquitetura, tecnologia e fornecedor permite que as decisões sejam tomadas rapidamente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

Ter políticas internas na seleção de recursos e arquitetura fornece padrões e diretrizes a serem seguidos ao fazer escolhas arquitetônicas. Essas diretrizes simplificam o processo de tomada de decisão ao escolher o serviço de nuvem certo e podem ajudar a melhorar a eficiência da performance. Implante a workload usando políticas ou arquiteturas de referência. Integre os serviços

à implantação na nuvem e, depois, use testes de desempenho para verificar se você pode continuar a atender aos seus requisitos de desempenho.

### Etapas da implementação

- Entenda claramente os requisitos de sua workload na nuvem.
- Analise as políticas internas e externas para identificar as mais relevantes.
- Use as arquiteturas de referência apropriadas fornecidas pela AWS ou as práticas recomendadas do seu setor.
- Crie um continuum que consiste em políticas, padrões, arquiteturas de referência e diretrizes prescritivas para situações comuns. Isso permite que suas equipes ajam mais rapidamente. Adapte os ativos para sua vertical, se aplicável.
- Valide essas políticas e arquiteturas de referência para sua workload em ambientes de sandbox.
- Atualize-se com relação aos padrões do setor e atualizações da AWS para garantir que suas políticas e arquiteturas de referência ajudem a otimizar sua workload na nuvem.

## Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS

Vídeos relacionados:

- This is my Architecture

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

# PERF01-BP06 Use testes comparativos para orientar decisões de arquitetura

Compare o desempenho de uma workload existente para entender seu desempenho na nuvem e orientar decisões de arquitetura com base nesses dados.

Antipadrões comuns:

- Você depende de testes comparativos comuns que não são indicativos das características da workload.
- Você conta com o feedback e as percepções de clientes como seu único teste comparativo.

Benefícios de estabelecer esta prática recomendada: Os testes comparativos da implementação atual permitem medir a melhoria da performance.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

Use testes comparativos com testes sintéticos para avaliar a performance dos componentes da workload. O benchmarking é usado na avaliação da tecnologia para um componente específico e geralmente é mais simples de configurar do que testes de carga. Muitas vezes o benchmarking é usado no início de um novo projeto, quando ainda não há uma solução completa para o teste de carga.

É possível criar os próprios testes comparativos personalizados ou usar um teste padrão do setor, como o TPC-DS, para comparar as workloads. Os benchmarks do setor são úteis ao comparar ambientes. Já os benchmarks personalizados são úteis para direcionar a tipos específicos de operações que você espera realizar em sua arquitetura.

Ao realizar testes comparativos, é importante “preaquecer” o ambiente de teste para obter resultados válidos. Execute o mesmo teste comparativo várias vezes para verificar a captura de qualquer variação ao longo do tempo.

Como normalmente é mais rápido executar testes comparativos do que testes de carga, eles podem ser usados mais cedo no pipeline de implantação e fornecer um feedback mais rápido sobre desvios de performance. Ao avaliar uma alteração significativa em um componente ou serviço, um benchmark pode ser uma maneira rápida de verificar se é possível justificar a iniciativa para

concretizar a alteração. O uso de testes comparativos em conjunto com testes de carga é importante porque o teste de carga informa como é a performance da workload em produção.

### Etapas da implementação

- Defina as métricas (como utilização da CPU, latência ou throughput) para avaliar o desempenho da workload.
- Identifique e configure uma ferramenta de testes comparativos adequada à workload. Você pode usar serviços da AWS (como o Amazon CloudWatch) ou uma ferramenta de terceiros compatível com a workload.
- Execute testes comparativos e monitore as métricas durante o teste.
- Analise e documente os resultados do teste comparativo para identificar gargalos e problemas.
- Use os resultados do teste para tomar decisões de arquitetura e ajustar a workload. Isso pode incluir a mudança de serviços ou a adoção de novos recursos.
- Teste novamente a workload após o ajuste.

## Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics

Vídeos relacionados:

- This is my Architecture
- Optimize applications through Amazon CloudWatch RUM
- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS
- Distributed Load Tests
- Measure page load time with Amazon CloudWatch Synthetics (Medição do tempo de carga da página com o Amazon CloudWatch Synthetics)
- Amazon CloudWatch RUM Web Client (Cliente da web do Amazon CloudWatch RUM)

# PERF01-BP07 Use uma abordagem baseada em dados para escolhas de arquitetura

Defina uma abordagem clara e baseada em dados para escolhas de arquitetura a fim de verificar se os serviços e configurações de nuvem corretos são usados para atender às suas necessidades comerciais específicas.

Antipadrões comuns:

- Você pressupõe que sua arquitetura atual é estática e não deve ser atualizada ao longo do tempo.
- Suas escolhas de arquitetura são baseadas em suposições.
- Você apresenta alterações de arquitetura ao longo do tempo sem justificativa.

Benefícios de estabelecer esta prática recomendada: Ao ter uma abordagem bem definida para fazer escolhas de arquitetura, você usa dados para influenciar o projeto das workloads e tomar decisões conscientes ao longo do tempo.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

Use a experiência interna e o conhecimento da nuvem ou de recursos externos, como casos de uso publicados ou whitepapers, para escolher recursos e serviços em sua arquitetura. Você deve ter um processo bem definido que incentive a experimentação e os testes comparativos com os serviços que podem ser usados em suas workloads.

Os atrasos de workloads críticas devem consistir não apenas em histórias de usuários que venham a oferecer funcionalidades relevantes para empresas e usuários, mas também em histórias técnicas que formem uma base de arquitetura para as workloads. Essa base é formada por novos avanços

em tecnologia e novos serviços e os adota com base em dados e justificativas adequadas. Isso verifica se a arquitetura permanece preparada para o futuro e não fica estagnada.

### Etapas da implementação

- Interaja com as principais partes interessadas para definir os requisitos das workloads, incluindo considerações de desempenho, disponibilidade e custo. Considere fatores como o número de usuários e o padrão de uso das workloads.
- Crie uma base de arquitetura ou uma lista de pendências de tecnologia que seja priorizada junto com a lista de pendências funcional.
- Avalie diferentes serviços em nuvem (para obter mais detalhes, consulte PERF01-BP01 Conheça e compreenda os serviços e recursos de nuvem disponíveis).
- Explore diferentes padrões de arquitetura, como microsserviços ou tecnologia sem servidor, que atendem aos requisitos de performance (para obter mais detalhes, consulte PERF01-BP02 Use a orientação de seu provedor de nuvem ou de um parceiro apropriado para aprender sobre padrões de arquitetura e práticas recomendadas).

- Consulte outras equipes, diagramas de arquitetura e recursos, como arquitetos de soluções da AWS, Centro de Arquitetura da AWSe AWS Partner Network, para ajudar você a escolher a arquitetura certa para sua workload.

- Defina métricas de desempenho, como produtividade e tempo de resposta, que podem ajudar você a avaliar o desempenho das workloads.
- Experimente e use métricas definidas para validar o desempenho da arquitetura selecionada.
- Monitore e faça ajustes contínuos conforme necessário para manter o desempenho ideal da arquitetura.
- Documente a arquitetura e as decisões selecionadas como referência para futuras atualizações e aprendizados.
- Revise e atualize constantemente a abordagem para seleção de arquitetura com base em aprendizados, novas tecnologias e métricas. Esses parâmetros podem indicar que é necessário mudar ou que há algum problema na abordagem atual.

## Recursos

Documentos relacionados:

- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS

Vídeos relacionados:

- This is my Architecture

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

# Computação e hardware

A opção ideal de computação para uma workload específica pode variar de acordo com o design, os padrões de uso e as definições de configuração da aplicação. As arquiteturas podem usar diferentes opções de computação para vários componentes e permitir diferentes recursos para aprimorar a performance. A seleção da opção de computação incorreta para uma arquitetura pode levar a uma menor eficiência de performance.

Essa área de foco compartilha orientações e práticas recomendadas sobre como identificar e otimizar as opções de computação para eficiência de desempenho na nuvem.

Práticas recomendadas

- PERF02-BP01 Selecione as melhores opções de computação para as workloads
- PERF02-BP02 Entenda a configuração e os recursos de computação disponíveis
- PERF02-BP03 Colete métricas relacionadas à computação
- PERF02-BP04 Configure e dimensione corretamente os recursos de computação
- PERF02-BP05 Dimensione recursos de computação dinamicamente
- PERF02-BP06 Use optimized hardware-based compute accelerators

## PERF02-BP01 Selecione as melhores opções de computação para as workloads

Selecionar a opção de computação mais adequada para suas workloads permite que você melhore o desempenho, reduza os custos desnecessários de infraestrutura e reduza os esforços operacionais necessários para mantê-las.

Antipadrões comuns:

- É usada a mesma opção de computação utilizada on-premises.
- Você não tem conhecimento das opções, dos atributos e das soluções de computação em nuvem e de como essas soluções podem melhorar a performance computacional.
- É provisionada em excesso uma opção de computação existente para atender aos requisitos de ajuste de escala ou performance quando uma opção alternativa de computação se alinharia às características da workload com mais precisão.

Benefícios de estabelecer esta prática recomendada: Ao identificar os requisitos de computação e avaliar as opções disponíveis, você pode tornar a workload mais eficiente em termos de recursos.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

Para otimizar as workloads na nuvem quanto à eficiência de desempenho, é importante selecionar as opções de computação mais apropriadas para seu caso de uso e requisitos de desempenho. A AWS fornece uma variedade de opções de computação que atendem a diferentes workloads na nuvem. Por exemplo, você pode usar o Amazon EC2 para iniciar e gerenciar servidores virtuais, o AWS Lambda para executar código sem precisar provisionar nem gerenciar servidores, o Amazon ECS ou Amazon EKS para executar e gerenciar contêineres ou AWS Batch para processar grandes volumes de dados em paralelo. Com base em sua escala e necessidades de computação, você deve escolher e configurar a solução ideal para sua situação. Você também pode considerar o uso de vários tipos de soluções de computação em uma única workload, pois cada uma tem suas próprias vantagens e desvantagens.

As etapas a seguir orientam você na seleção das opções de computação certas para atender às características da workload e aos requisitos de desempenho.

## Etapas da implementação

1. Entenda os requisitos de computação das workloads. Os principais requisitos a serem considerados incluem necessidades de processamento, padrões de tráfego, padrões de acesso a dados, necessidades de ajuste de escala e requisitos de latência.
2. Saiba mais sobre as diferentes opções de computação disponíveis para a workload na AWS (conforme descrito em PERF01-BP01 Conheça e compreenda os serviços e recursos de nuvem disponíveis. Veja algumas das principais opções de computação da AWS, as características e casos de uso comuns:

| Serviço da AWS | Características principais | Casos de uso comum |
| --- | --- | --- |
| Amazon Elastic Compute Cloud (Amazon EC2) | Tem opção dedicada para hardware, requisitos de licença, grande seleção de diferentes famílias de instâncias, tipos de | Migrações do tipo mover sem alterações (lift-and-shift), aplicações monolíticas, ambientes híbridos, aplicaçõe s empresariais |

| Serviço da AWS | Características principais | Casos de uso comum |
| --- | --- | --- |
| | processadores e acelerado res de computação. | |
| Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) | Fácil implantação, ambientes consistentes, escaláveis | Microsserviços, ambientes híbridos |
| AWS Lambda | Computação com tecnologi a sem servidor Serviço que executa código em resposta a eventos e gerencia automaticamente os recursos computacionais subjacentes. | Microsserviços, aplicações orientadas a eventos |
| AWS Batch | Provisiona e escala de forma eficiente e dinâmica. Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)e AWS Fargate Recursos de computação, com a opção de usar instâncias sob demanda ou spot com base nos requisitos de trabalho. | HPC, treine modelos de ML. |
| Amazon Lightsail | Aplicação Linux e Windows pré-configurada para executar pequenas workloads . | Aplicações web simples, site personalizado. |

3. Avalie o custo (como cobrança por hora ou transferência de dados) e as despesas gerais de gerenciamento (como aplicação de patches e ajuste de escala) associados a cada opção de computação.

4. Faça experimentos e análises comparativas em um ambiente que não seja de produção para identificar qual opção de computação pode melhor atender às necessidades da workload.
5. Depois de experimentar e identificar sua nova solução de computação, planeje a migração e valide as métricas de desempenho.
6. Use ferramentas de monitoramento da AWS, como Amazon CloudWatch e serviços de otimização, como AWS Compute Optimizer para otimizar continuamente os recursos de computação com base em padrões de uso do mundo real.

## Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Amazon EC2 Instance Types
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Amazon ECS Containers: Amazon ECS Container Instances
- Funções: configuração de função do Lambda
- Prescriptive Guidance for Containers
- Prescriptive Guidance for Serverless

Vídeos relacionados:

- How to choose compute option for startups (Como escolher uma opção de computação para startups)
- Optimize performance and cost for your AWS compute
- Amazon EC2 foundations
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Deploy ML models for inference at high performance and low cost
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2

Exemplos relacionados:

- Migrating the Web application to containers

- Run a Serverless Hello World

# PERF02-BP02 Entenda a configuração e os recursos de computação disponíveis

Entenda as opções de configuração e os recursos disponíveis para seu serviço de computação a fim de ajudar a provisionar a quantidade certa de recursos e melhorar a eficiência do desempenho.

Antipadrões comuns:

- Você não avalia as opções de computação ou as famílias de instâncias disponíveis em relação às características da workload.
- Você provisiona recursos de computação em excesso para atender aos requisitos de pico de demanda.

Benefícios de estabelecer esta prática recomendada: familiarizar-se com os atributos e as configurações de computação da AWS a fim de poder usar uma solução de computação otimizada para atender às características e às necessidades da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

Cada solução de computação tem configurações e recursos exclusivos disponíveis para suportar diferentes características e requisitos das workloads. Saiba como essas opções complementam sua workload e determine quais opções de configuração são melhores para sua aplicação. Exemplos dessas opções são famílias de instâncias, tamanhos, recursos (GPU, E/S), expansão, tempos limite, tamanhos de função, instâncias de contêineres e simultaneidade. Se a workload estiver usando a mesma opção de computação há mais de quatro semanas, e se a previsão for de que as características permanecerão as mesmas no futuro, você poderá usar o AWS Compute Optimizer para descobrir se a opção de computação atual é adequada para as workloads do ponto de vista da CPU e da memória.

## Etapas da implementação

1. Entenda os requisitos da workload (como necessidade de CPU, memória e latência).

2. Analise a documentação e as práticas recomendadas da AWS para saber mais sobre as opções de configuração indicadas que podem ajudar a melhorar a performance da computação. Aqui estão algumas das principais opções de configuração a serem consideradas:

| Opção de configuração | Exemplos |
| --- | --- |
| Tipo de instância | • As instâncias otimizadas para computação são ideais para workloads que exigem uma proporção maior de vCPU/memória.<br>• As instâncias otimizadas para memória entregam grandes quantidades de memória para oferecer compatibilidade com as workloads com uso intenso de memória.<br>• As instâncias otimizadas para armazenamento são projetadas para workloads que exigem alta leitura sequencial e acesso de gravação (IOPS) no armazenamento local. |
| Modelo de definição de preço | • Instâncias sob demanda permitem usar a capacidade de computação pela hora ou segundo sem uma confirmação de longo prazo. Essas instâncias são ideais para expansões acima das necessidades de desempenho da linha de base.<br>• Savings Plans oferecem economias significativas em relação às instâncias sob demanda em troca do compromisso de usar uma quantidade específica de potência computacional por um período de um ou três anos.<br>• instâncias spot. permitem que você aproveite a capacidade da instância não utilizada com um desconto para as workloads sem estado e tolerantes a falhas. |

| Opção de configuração | Exemplos |
| --- | --- |
| Auto Scaling | Use o Auto Scaling configuração para combinar recursos computacionais com padrões de tráfego. |
| Dimensionamento | • Use Compute Optimizer para obter uma recomendação de machine learning sobre a configuração de computação que correspon de melhor às características da computaçã o.<br>• Use AWS Lambda Power Tuning para selecionar a melhor configuração para a função do Lambda. |
| Aceleradores de computação baseados em hardware | • As instâncias com computação acelerada executam funções como processamento gráfico ou correspondência de padrões de dados com mais eficiência do que as alternativas baseadas em CPU.<br>• Para workloads de machine learning, utilize hardware específico para sua workload, como AWS Trainium, AWS Inferentiae o Amazon EC2 DL1 |

## Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Amazon EC2 Instance Types
- Processor State Control for Your Amazon EC2 Instance
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Amazon ECS Containers: Amazon ECS Container Instances
- Funções: configuração de função do Lambda

Vídeos relacionados:

- Amazon EC2 foundations
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Optimize performance and cost for your AWS compute

Exemplos relacionados:

- Rightsizing with Compute Optimizer and Memory utilization enabled
- AWS Compute Optimizer Demo code

# PERF02-BP03 Colete métricas relacionadas à computação

Registre e acompanhe métricas relacionadas à computação para entender melhor o desempenho de seus recursos e melhorar seu desempenho e utilização.

Antipadrões comuns:

- Você só usa a pesquisa manual de arquivos de log para métricas.
- Você só usa as métricas padrão registradas pelo software de monitoramento.
- Você só revisa as métricas quando há um problema.

Benefícios de estabelecer esta prática recomendada: A coleta de métricas relacionadas à performance ajudará você a alinhar a performance da aplicação aos requisitos empresariais para garantir que você atenda às necessidades da workload. Isso também pode ajudar a melhorar constantemente o desempenho e a utilização dos recursos na workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

As workloads na nuvem podem gerar grandes volumes de dados, como métricas, logs e eventos. Na Nuvem AWS, coletar métricas é uma etapa essencial para melhorar a segurança, a eficiência de custos, a performance e a sustentabilidade. A AWS oferece uma ampla variedade de métricas relacionadas à performance usando serviços de monitoramento, por exemplo, o Amazon CloudWatch para fornecer informações valiosas. Métricas como utilização de CPU, utilização de

memória, E/S de disco e entrada e saída da rede podem fornecer informações sobre os níveis de utilização ou gargalos de desempenho. Use essas métricas como parte de uma abordagem impulsionada por dados para ajustar e otimizar ativamente os recursos de sua carga de trabalho. Em um caso ideal, você deve coletar todas as métricas relacionadas aos recursos de computação em uma única plataforma com políticas de retenção implementadas para apoiar as metas operacionais e de custo.

## Etapas da implementação

1. Identifique quais métricas relacionadas ao desempenho são relevantes para a workload. Você deve coletar métricas sobre a utilização de recursos e a forma como a workload na nuvem está operando (como tempo de resposta e throughput).
    a. Métricas padrão do Amazon EC2
    b. Métricas padrão do Amazon ECS
    c. Métricas padrão do Amazon EKS
    d. Métricas padrão do Lambda
    e. Métricas de memória e disco do Amazon EC2
2. Escolha e configure a solução certa de registro e monitoramento para a workload.
    a. Observabilidade nativa da AWS
    b. AWS Distro para OpenTelemetry
    c. Amazon Managed Service for Prometheus
3. Defina o filtro e a agregação necessários para as métricas com base nos requisitos da workload.
    a. Quantify custom application metrics with Amazon CloudWatch Logs and metric filters
    b. Collect custom metrics with Amazon CloudWatch strategic tagging
4. Configure políticas de retenção de dados para que as métricas correspondam às metas operacionais e de segurança.
    a. Retenção de dados padrão para métricas do CloudWatch
    b. Retenção de dados padrão para o CloudWatch Logs
5. Se necessário, crie alarmes e notificações para as métricas a fim de ajudar a reagir proativamente a problemas relacionados à performance.
    a. Create alarms for custom metrics using Amazon CloudWatch anomaly detection
    b. Create metrics and alarms for specific web pages with Amazon CloudWatch RUM
6. Use a automação para implantar os agentes de agregação de métricas e logs.

a. AWS Systems Manager Automation
b. OpenTelemetry Collector

## Recursos

Documentos relacionados:

- Documentação do Amazon CloudWatch
- Collect metrics and logs from Amazon EC2 instances and on-premises servers with the CloudWatch Agent
- Acessar o Amazon CloudWatch Logs para o AWS Lambda
- Using CloudWatch Logs with container instances
- Publicar métricas personalizadas
- AWS Answers: Centralized Logging
- AWS Services That Publish CloudWatch Metrics
- Monitoring Amazon EKS on AWS Fargate

Vídeos relacionados:

- Application Performance Management on AWS

Exemplos relacionados:

- Level 100: Monitoring with CloudWatch Dashboards
- Level 100: Monitoring Windows EC2 instance with CloudWatch Dashboards
- Level 100: Monitoring an Amazon Linux EC2 instance with CloudWatch Dashboards

# PERF02-BP04 Configure e dimensione corretamente os recursos de computação

Configure e dimensione corretamente os recursos de computação para atender aos requisitos de desempenho das workloads e evitar que recursos sejam subutilizados ou usados em excesso.

Antipadrões comuns:

- Ignorar os requisitos de performance das workloads, o que ocasiona recursos computacionais superprovisionados ou subprovisionados.
- Você escolhe somente a maior ou a menor instância disponível para todas as workloads.
- Você usa apenas uma família de instâncias para facilitar o gerenciamento.
- Você ignora as recomendações de AWS Cost Explorer ou Compute Optimizer para o dimensionamento correto.
- Você não reavalia a workload quanto à adequação dos novos tipos de instância.
- Você certifica apenas um pequeno número de configurações de instâncias para sua organização.

Benefícios de estabelecer esta prática recomendada: o dimensionamento correto dos recursos computacionais garante a operação ideal na nuvem, evitando o provisionamento excessivo e o subprovisionamento de recursos. O dimensionamento adequado dos recursos de computação normalmente resulta em melhor desempenho e melhor experiência do cliente, além de reduzir custos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

O dimensionamento correto permite que as organizações operem a infraestrutura de nuvem de forma eficiente e econômica, ao mesmo tempo em que atendem às suas necessidades comerciais. O provisionamento excessivo de recursos de nuvem pode gerar custos extras, enquanto o subprovisionamento pode ocasionar performance ruim e uma experiência negativa para o cliente. A AWS oferece ferramentas, como o AWS Compute Optimizer e o AWS Trusted Advisor , que usam dados históricos com o objetivo de fornecer recomendações para dimensionar corretamente os recursos computacionais.

### Etapas da implementação

- Escolha um tipo de instância que melhor atenda às suas necessidades:
  - Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?
  - Seleção de tipo de instância baseada em atributos para frota do Amazon EC2
  - Create an Auto Scaling group using attribute-based instance type selection
  - Optimizing your Kubernetes compute costs with Karpenter consolidation (Otimizar seus custos de computação do Kubernetes com a consolidação do Karpenter)

- Analise as várias características de performance de sua carga de trabalho e como elas se relacionam a uso de memória, rede e CPU. Use esses dados para escolher os recursos que melhor correspondam ao perfil e às metas de desempenho da workload.
- Monitore o uso de recursos usando ferramentas de monitoramento da AWS, como o Amazon CloudWatch.
- Selecione a configuração correta para os recursos computacionais.
  - Para workloads efêmeras, avalie métricas do Amazon CloudWatch para instâncias , como `CPUUtilization` para identificar se a instância está subutilizada ou superutilizada.
  - Para workloads estáveis, verifique as ferramentas de dimensionamento correto da AWS, como AWS Compute Optimizer e AWS Trusted Advisor em intervalos regulares, para identificar oportunidades de otimizar e dimensionar corretamente o recurso de computação.
    - Laboratório do Well-Architected: Recomendações de dimensionamento correto
    - Laboratório do Well-Architected: Dimensionamento correto com o Compute Optimizer
- Teste as alterações na configuração em um ambiente que não seja de produção antes de implementá-las em um ambiente ativo.
- Reavalie constantemente novas ofertas de computação e as compare com as necessidades da workload.

## Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Amazon EC2 Instance Types
- Amazon ECS Containers: Amazon ECS Container Instances
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Funções: configuração de função do Lambda
- Processor State Control for Your Amazon EC2 Instance

Vídeos relacionados:

- Amazon EC2 foundations
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2
- Deploy ML models for inference at high performance and low cost

- Optimize performance and cost for your AWS compute
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Como simplificar o processamento de dados para aprimorar a inovação com ferramentas de tecnologia sem servidor

Exemplos relacionados:

- Rightsizing with Compute Optimizer and Memory utilization enabled
- AWS Compute Optimizer Demo code

# PERF02-BP05 Dimensione recursos de computação dinamicamente

Use a elasticidade da nuvem para aumentar ou diminuir os recursos de computação dinamicamente a fim de atender às suas necessidades e evitar provisionamento excessivo ou insuficiente da capacidade para a workload.

Antipadrões comuns:

- Você reage a alarmes aumentando a capacidade manualmente.
- Você usa as mesmas diretrizes de dimensionamento (geralmente infraestrutura estática) do ambiente on-premises.
- Você deixa a capacidade aumentada após um evento de escalabilidade, em vez de reduzir novamente.

Benefícios de estabelecer esta prática recomendada: Configurar e testar a elasticidade dos recursos computacionais pode ajudar você a economizar dinheiro, manter os benchmarks de performance e melhorar a confiabilidade à medida que o tráfego muda.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

A AWS oferece a flexibilidade de aumentar ou diminuir seus recursos dinamicamente por meio de uma variedade de mecanismos de ajuste de escala a fim de atender às mudanças na demanda.

Combinado com métricas relacionadas à computação, um ajuste de escala dinâmico permite que as workloads respondam automaticamente às mudanças e usem o conjunto ideal de recursos computacionais para atingir sua meta.

Você pode usar diversas abordagens diferentes para corresponder a oferta de recursos com a demanda.

- Abordagem de monitoramento de meta: monitore a métrica de ajuste de escala e aumente ou diminua automaticamente a capacidade conforme necessário.
- Ajuste de escala preditivo: reduza a escala horizontalmente em antecipação às tendências diárias e semanais.
- Abordagem baseada em cronograma: defina seu próprio cronograma de escalabilidade de acordo com as mudanças de carga previsíveis.
- Escalabilidade de serviços: escolha serviços (como de tecnologia sem servidor) que sejam escalados automaticamente de acordo com o projeto.

É necessário garantir que as implantações de carga de trabalho possam lidar com eventos de expansão e redução da escala.

## Etapas da implementação

- Instâncias, contêineres e funções de computação oferecem mecanismos para elasticidade, seja em combinação com o ajuste de escala automático ou como um recurso do serviço. Veja alguns exemplos de mecanismos de ajuste de escala automático:

| Mecanismo de ajuste de escala automático | Onde usar |
| --- | --- |
| Amazon EC2 Auto Scaling | Para garantir que você tenha o número correto de instâncias do Amazon EC2 disponíveis para lidar com a carga do usuário para a aplicação. |
| Application Auto Scaling | Para escalar automaticamente os recursos para serviços individuais da AWS além do Amazon EC2, como funções do AWS Lambda ou os serviços Amazon Elastic Container Service (Amazon ECS) . |

| Mecanismo de ajuste de escala automático | Onde usar |
| --- | --- |
| Kubernetes Cluster Autoscaler/Karpenter | Para escalar automaticamente os clusters do Kubernetes. |

- O ajuste de escala geralmente é discutido em relação a serviços de computação, como instâncias do Amazon EC2 ou funções do AWS Lambda. Não se esqueça de pensar também na configuração de serviços não computacionais, como AWS Glue para atender à demanda.
- Verifique se as métricas de ajuste de escala correspondem às características da workload que está sendo implantada. Se você estiver implantando uma aplicação de transcodificação de vídeo, espera-se que a utilização da CPU seja de 100%, e essa não deve ser sua métrica principal. Em vez disso, use a profundidade da fila de trabalhos de transcodificação. Você pode usar uma métrica personalizada para a política de escalabilidade, se necessário. Para escolher as métricas certas, considere a seguinte orientação para o Amazon EC2:
  - A métrica deve ser uma métrica de utilização válida e descrever o quanto uma instância está ocupada.
  - O valor da métrica deve aumentar ou diminuir proporcionalmente com o número de instâncias no grupo do Auto Scaling.
- Use a escalabilidade dinâmica em vez de escalabilidade manual para seu grupo do Auto Scaling. Também recomendamos que você use políticas de escalabilidade de monitoramento do objetivo em sua escalabilidade dinâmica.
- Verifique se as implantações da workload podem lidar com os dois eventos de ajuste de escala (aumento e redução). Por exemplo, você pode usar o histórico de atividades para verificar uma atividade de escalabilidade para um grupo do Auto Scaling.
- Avalie sua workload com relação a padrões previsíveis e, ao antecipar alterações previstas e planejadas na demanda, escale proativamente. Com a escalabilidade preditiva, é possível eliminar a necessidade de superprovisionar a capacidade. Para obter mais detalhes, consulte Ajuste de escala com o Amazon EC2 Auto Scaling.

## Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Amazon EC2 Instance Types

- Amazon ECS Containers: Amazon ECS Container Instances
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Funções: configuração de função do Lambda
- Processor State Control for Your Amazon EC2 Instance
- Deep Dive on Amazon ECS Cluster Auto Scaling
- Introducing Karpenter – An Open-Source High-Performance Kubernetes Cluster Autoscaler (Apresentação do Karpenter: um dimensionador automático de clusters do Kubernetes de código aberto e alto desempenho)

Vídeos relacionados:

- Amazon EC2 foundations
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2
- Optimize performance and cost for your AWS compute
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Criar um ambiente de computação eficiente em termos de custo, energia e recursos

Exemplos relacionados:

- Amazon EC2 Auto Scaling Group Examples
- Implement Autoscaling with Karpenter

## PERF02-BP06 Use optimized hardware-based compute accelerators

Use aceleradores de hardware para executar determinadas funções com mais eficiência do que as alternativas baseadas em CPU.

Antipadrões comuns:

- Em sua workload, você não compara uma instância de uso geral com uma instância criada para um propósito específico que possa oferecer maior desempenho e menor custo.
- Você está usando aceleradores de computação baseados em hardware para tarefas que podem ser mais eficientes usando alternativas baseadas em CPU.

- Você não está monitorando o uso da GPU.

Benefícios de estabelecer esta prática recomendada: Ao usar aceleradores baseados em hardware, como unidades de processamento gráfico (GPUs) e matrizes de portas programáveis em campo (FPGAs), você pode executar determinadas funções de processamento com mais eficiência.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

As instâncias com computação acelerada fornecem acesso a aceleradores de computação baseados em hardware, como GPUs e FPGAs. Esses aceleradores de hardware executam certas funções, como processamento gráfico ou correspondência de padrões de dados, com mais eficiência do que alternativas baseadas em CPU. Muitas workloads aceleradas, como renderização, transcodificação e machine learning, são altamente variáveis em termos de uso de recursos. Execute esse hardware apenas pelo tempo necessário e desative-as com automação quando não precisar mais delas para melhorar a eficiência geral do desempenho.

### Etapas da implementação

- Identifique quais instâncias com computação acelerada podem atender aos seus requisitos.
- Para workloads de machine learning, utilize hardware específico para sua workload, como AWS Trainium, AWS Inferentiae o Amazon EC2 DL1. Instâncias do AWS Inferentia, como instâncias Inf2, oferecem até 50% melhor performance/watt em relação a instâncias comparáveis do Amazon EC2.
- Colete métricas de uso para as instâncias com computação acelerada. Por exemplo, você pode usar o agente do CloudWatch para coletar métricas como `utilization_gpu` e `utilization_memory` para suas GPUs, conforme mostrado em Colete métricas da GPU NVIDIA com o Amazon CloudWatch.
- Otimize o código, a operação de rede e as configurações dos aceleradores de hardware para garantir que o hardware subjacente seja totalmente utilizado.
  - Otimizar as configurações da GPU
  - Monitoramento e otimização de GPU no Deep Learning AMI
  - Otimização de E/S para ajuste de desempenho de GPU de treinamento de aprendizado profundo no Amazon SageMaker
- Use as mais recentes bibliotecas de alto desempenho e drivers de GPU.

- Use automação para liberar instâncias de GPU quando não estiverem em uso.

## Recursos

Documentos relacionados:

- Instâncias de GPU
- Instâncias com AWS Trainium
- Instâncias com o AWS Inferentia
- Let's Architect! Arquitetura com chips e aceleradores personalizados

- Computação acelerada
- Instâncias VT1 do Amazon EC2
- Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?
- Escolha o melhor acelerador de IA e compilação de modelo para inferência de visão computacional com o Amazon SageMaker

Vídeos relacionados:

- How to select Amazon EC2 GPU instances for deep learning (Como selecionar instâncias de GPU do Amazon EC2 para aprendizado profundo)
- Deploying Cost-Effective Deep Learning Inference (Implantação de inferência de aprendizado profundo econômico)

# Gerenciamento de dados

A solução de gerenciamento de dados ideal para um sistema específico varia conforme o tipo de dados (bloco, arquivo ou objeto), os padrões de acesso (aleatório ou sequencial), o throughput necessário, a frequência de acesso (online, offline, arquivamento), a frequência de atualização (WORM, dinâmica) e as restrições de disponibilidade e durabilidade. As workloads do Well-Architected usam datastores específicos que permitem que recursos diferentes melhorem o desempenho.

Essa área de foco compartilha orientações e práticas recomendadas para otimizar o armazenamento de dados, os padrões de movimentação e acesso, e a eficiência do desempenho dos armazenamentos de dados.

Práticas recomendadas

- PERF03-BP01 Use um armazenamento de dados específico que melhor atenda aos seus requisitos de acesso e armazenamento de dados
- PERF03-BP02 Avalie as opções de configuração disponíveis para o datastore
- PERF03-BP03 Colete e registre métricas de desempenho do datastore
- PERF03-BP04 Implemente estratégias para melhorar o desempenho da consulta no datastore
- PERF03-BP05 Implementar padrões de acesso a dados que utilizem cache

## PERF03-BP01 Use um armazenamento de dados específico que melhor atenda aos seus requisitos de acesso e armazenamento de dados

Entenda as características dos dados (como possibilidade de compartilhamento, tamanho, tamanho do cache, padrões de acesso, latência, throughput e persistência dos dados) a fim de selecionar os datastores com propósito específico (armazenamento ou banco de dados) para sua workload.

Antipadrões comuns:

- Fixar-se em um único datastore porque há experiência e conhecimento internos de um tipo específico de solução de banco de dados.
- Você pressupõe que todas as workloads têm requisitos de acesso e armazenamento de dados semelhantes.

- Você não implementou um catálogo de dados para criar um inventário de seus ativos de dados.

Benefícios de estabelecer esta prática recomendada: Entender as características e os requisitos de dados permite que você determine a tecnologia de armazenamento mais eficiente e com melhor performance, adequada às necessidades da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

Ao selecionar e implementar o armazenamento de dados, certifique-se de que as características de consulta, ajuste de escala e armazenamento atendam aos requisitos de dados da workload. A AWS fornece várias tecnologias de armazenamento de dados e banco de dados, incluindo armazenamento em blocos, armazenamento de objetos, armazenamento de streaming, sistema de arquivos, bancos de dados relacionais, de chave-valor, de documentos, na memória, de grafos, de séries temporais e ledger. Cada solução de gerenciamento de dados tem opções e configurações disponíveis para compatibilidade com seus casos de uso e modelos de dados. Ao compreender as características e os requisitos dos dados, você pode se separar da tecnologia de armazenamento monolítico e das abordagens restritivas e únicas para se concentrar no gerenciamento adequado dos dados.

### Etapas da implementação

- Realize um inventário dos vários tipos de dados que existem na workload.
- Entenda e documente as características e os requisitos dos dados, incluindo:
  - Tipo de dados (não estruturados, semiestruturados, relacionais)
  - Volume e crescimento de dados
  - Durabilidade dos dados: persistentes, efêmeros, transitórios
  - Requisitos de ACID (atomicidade, consistência, isolamento, durabilidade)
  - Padrões de acesso a dados (com muita leitura ou gravação)
  - Latência
  - Taxa de transferência
  - IOPS (operações de entrada/saída por segundo)
  - Período de retenção de dados
- Conheça os diferentes datastores disponíveis para a workload na AWS que podem atender às características dos dados (conforme descrito em PERF01-BP01 Conheça e compreenda os

serviços e recursos de nuvem disponíveis). Alguns exemplos de tecnologias de armazenamento da AWS e suas principais características incluem:

| Tipo | Serviços da AWS | Características principais |
| --- | --- | --- |
| Armazenamento de objetos | Amazon S3 | Escalabilidade ilimitada, alta disponibilidade e várias opções de acessibilidade. A transferência e o acesso a objetos dentro e fora do Amazon S3 podem usar um serviço, como Aceleração de Transferências ou Pontos de Acesso, para oferecer compatibilidade com o local, necessidades de segurança e padrões de acesso. |
| Armazenamento de arquivamento | Amazon S3 Glacier | Desenvolvido para arquivamento de dados. |
| Armazenamento de streaming | Amazon Kinesis<br><br>Amazon Managed Streaming for Apache Kafka (Amazon MSK) | Ingestão e armazenamento eficientes de dados de streaming. |
| Sistema de arquivos compartilhado | Amazon Elastic File System (Amazon EFS) | Sistema de arquivos montável que pode ser acessado por vários tipos de soluções de computação. |

| Tipo | Serviços da AWS | Características principais |
| --- | --- | --- |
| Sistema de arquivos compartilhado | Amazon FSx | Baseia-se nas soluções de computação mais recentes da AWS para oferecer compatibilidade com quatro sistemas de arquivos usados com frequência: NetApp ONTAP, OpenZFS, Windows File Server e Lustre. Amazon FSx latência, throughput e IOPS variam de acordo com o sistema de arquivos e devem ser consideradas ao selecionar o sistema de arquivos certo para as necessidades de sua workload. |
| O Armazenamento em bloco | Amazon Elastic Block Store (Amazon EBS) | Serviço de armazenamento de bloco escalável e de alta performance projetado para Amazon Elastic Compute Cloud (Amazon EC2). O Amazon EBS inclui armazenamento com base em SSD para workloads transacionais e de uso intenso de IOPS e armazenamento com base em HDD para workloads de uso intenso de throughput. |

| Tipo | Serviços da AWS | Características principais |
| --- | --- | --- |
| Banco de dados relacional | Amazon Aurora, o Amazon RDS, o Amazon Redshift. | Projetados para oferecer compatibilidade com transações ACID (atomicidade, consistência, isolamento, durabilidade) e manter a integridade referencial e uma forte consistência de dados. Muitas aplicações tradicionais, planejamento de recursos empresariais (ERP), gerenciamento de relacionamentos com o cliente (CRM) e comércio eletrônico usam bancos de dados relacionais para armazenar os dados. |
| Banco de dados de chave-valor | tabelas do Amazon DynamoDB | Otimizados para padrões de acesso comuns, normalmente visando armazenar e recuperar grandes volumes de dados. Aplicações web de alto tráfego, sistemas de comércio eletrônico e aplicações de jogos são os casos de uso habituais para bancos de dados de chave-valor. |

| Tipo | Serviços da AWS | Características principais |
| --- | --- | --- |
| Banco de dados de documentos | Amazon DocumentDB | Projetado para armazenar dados semiestruturados, como documentos do tipo JSON. Esses bancos de dados ajudam os desenvolvedores a criar e atualizar rapidamente aplicativos como gerenciamento de conteúdo, catálogos e perfis de usuário. |
| Banco de dados na memória | Amazon ElastiCache , Amazon MemoryDB for Redis | Usados para aplicações que exigem acesso em tempo real aos dados, latência mais baixa e throughput mais alto. É possível usar bancos de dados na memória para armazenamento em cache de aplicações, gerenciamento de sessões, tabelas de classificação de jogos, arquivo de atributos de ML de baixa latência, sistema de mensagens de microsserviços e um mecanismo de streaming de alto throughput. |

| Tipo | Serviços da AWS | Características principais |
| --- | --- | --- |
| Banco de dados de grafos | Amazon Neptune | Utilizado para aplicações que precisam navegar e consultar milhões de relacionamentos entre conjuntos de dados de grafos altamente conectados com latência de milissegundos em grande escala. Muitas empresas usam bancos de dados gráficos para detecção de fraudes, redes sociais e mecanismos de recomendação. |
| Banco de dados de séries temporais | Amazon Timestream | Utilizado para coletar, sintetizar e gerar com eficiência insights de dados que mudam ao longo do tempo. Aplicativos de IoT, DevOps e telemetria industrial podem utilizar bancos de dados de séries temporais. |
| Coluna ampla | Amazon Keyspaces (para Apache Cassandra) | Usa tabelas, linhas e colunas, mas ao contrário de um banco de dados relacional, os nomes e o formato das colunas podem variar de linha para linha na mesma tabela. Normalmente, você vê um repositório de coluna ampla em aplicativos industriais de alta escala para manutenção de equipamentos, gerenciamento de frotas e otimização de rotas. |

| Tipo | Serviços da AWS | Características principais |
| --- | --- | --- |
| Ledger | Amazon Quantum Ledger Database (Amazon QLDB) | Oferece uma autoridade centralizada e confiável para manter um registro escalável, imutável e criptograficamente verificável de transações para cada aplicação. Vemos os bancos de dados de livro-razão empregados em sistemas de registro, cadeia de suprimentos, inscrições e até mesmo transações bancárias. |

- Se você estiver criando uma plataforma de dados, utilize uma arquitetura de dados moderna na AWS para integrar data lake, data warehouse e armazenamentos de dados com propósito específico.
- As principais questões que você precisa considerar ao escolher um datastore para sua workload são as seguintes:

| Pergunta | Fatos a serem considerados |
| --- | --- |
| Como os dados são estruturados? | • Se os dados estiverem estruturados, pense em um armazenamento de objetos, como o Amazon S3 , ou um banco de dados NoSQL, como o Amazon DocumentDB<br>• Para dados de chave-valor, pense no DynamoDB, o Amazon ElastiCache for Redis ou Amazon MemoryDB for Redis |
| Qual nível de integridade referencial é necessário? | • Para restrições de chave externa, bancos de dados relacionais, como o Amazon RDS e o Aurora , podem oferecer esse nível de integridade. |

| Pergunta | Fatos a serem considerados |
| --- | --- |
| | • Normalmente, em um modelo de dados NoSQL, você desnormalizaria os dados em um único documento ou coleção de documentos para serem recuperados em uma única solicitação em vez de unir documentos ou tabelas de diferentes locais. |
| A conformidade com ACID (atomicidade, consistência, isolamento, durabilidade) é necessária? | • Se as propriedades ACID associadas aos bancos de dados relacionais forem necessárias, pense em um banco de dados relacional, como o Amazon RDS e o Aurora.<br>• Se for necessária uma consistência forte para banco de dados NoSQL, você pode usar leituras altamente consistentes com DynamoDB. |
| Como as necessidades de armazenamento serão alteradas ao longo do tempo? Como isso afeta a escalabilidade? | • Bancos de dados de tecnologia sem servidor como o DynamoDB e o Amazon Quantum Ledger Database (Amazon QLDB) serão escalados dinamicamente.<br>• Os bancos de dados relacionais têm limites superiores em armazenamento provisionado e devem ser particionados horizontalmente usando mecanismos, como fragmentação, quando atingem esses limites. |

| Pergunta | Fatos a serem considerados |
| --- | --- |
| Qual é a proporção de consultas de leitura em relação a consultas de gravação? O armazenamento em cache melhoraria a performance? | • Workloads de uso intenso de leitura podem se beneficiar de uma camada de armazenamento em cache, como o ElastiCache ou DAX se o banco de dados for o DynamoDB.<br>• As leituras também podem ser descarreg adas em réplicas de leitura com bancos de dados relacionais, como o Amazon RDS. |
| O armazenamento e a modificação (OLTP – Processamento de transações on-line) ou a recuperação e a geração de relatórios (OLAP – Processamento analítico on-line) têm uma prioridade mais alta? | • Para um processamento transacional de throughput alto de leitura no estado em que se encontra, considere um banco de dados NoSQL, como o DynamoDB.<br>• Para padrões de leitura complexos e de throughput alto (como junção) com consistência use o Amazon RDS.<br>• Para consultas de análise, pense em um banco de dados em colunas, como o Amazon Redshift , ou exporte os dados para o Amazon S3 e realize análises usando o Athena ou Amazon QuickSight. |

| Pergunta | Fatos a serem considerados |
| --- | --- |
| Qual nível de durabilidade os dados exigem? | • O Aurora replica automaticamente os dados entre três zonas de disponibilidade em uma região, o que significa que seus dados terão mais durabilidade com menos chance de serem perdidos.<br>• O DynamoDB é automaticamente replicado entre várias zonas de disponibilidade, fornecendo alta disponibilidade e durabilid ade aos dados.<br>• O Amazon S3 fornece 11 noves de durabilid ade. Muitos serviços de banco de dados, como o Amazon RDS e o DynamoDB, são compatíveis com a exportação de dados ao Amazon S3 para retenção de longo prazo e arquivamento. |
| Você quer se livrar de mecanismos de bancos de dados comerciais ou custos de licenças? | • Considere os mecanismos de código aberto, como o PostgreSQL e o MySQL no Amazon RDS ou no Aurora.<br>• Utilize o AWS Database Migration Service e o AWS Schema Conversion Tool para realizar migrações de mecanismos de bancos de dados comerciais para código aberto |
| Qual a expectativa operacional para o banco de dados? A mudança para serviços gerenciados é uma preocupação principal? | • Utilizar o Amazon RDS em vez do Amazon EC2 e o DynamoDB ou o Amazon DocumentDB em vez de um host automátic o de um banco de dados NoSQL pode reduzir a sobrecarga operacional. |

| Pergunta | Fatos a serem considerados |
| --- | --- |
| Como o banco de dados é acessado atualmente? É acessado apenas por aplicação ou há usuários de inteligência de negócios (BI) e outras aplicações prontas para uso conectadas? | • Se você tiver dependências de ferramentas externas, poderá ser necessário manter a compatibilidade com os bancos de dados com os quais elas são compatíveis. O Amazon RDS é totalmente compatível com as diferentes versões de mecanismo aos quais oferece suporte, incluindo o Microsoft SQL Server, o Oracle, o MySQL e o PostgreSQL. |

- Faça experimentos e testes comparativos em um ambiente que não seja de produção para identificar qual datastore pode atender às necessidades da workload.

## Recursos

Documentos relacionados:

- Tipos de volume do Amazon EBS
- Armazenamento do Amazon EC2
- Amazon EFS: performance do Amazon EFS
- Performance do Amazon FSx for Lustre
- Performance do Amazon FSx for Windows File Server
- Documentação do Amazon S3 Glacier: S3 Glacier
- Amazon S3: considerações sobre performance e taxa de solicitação
- Armazenamento na nuvem com a AWS
- Características de E/S do Amazon EBS
- Bancos de dados em nuvem com a AWS
- Armazenamento em cache de banco de dados da AWS
- DynamoDB Accelerator
- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- 10 melhores dicas de desempenho do Amazon Athena

- Práticas recomendadas do Amazon Redshift Spectrum
- Melhores práticas do Amazon DynamoDB
- Choose between Amazon EC2 and Amazon RDS
- Melhores práticas para a implementação do Amazon ElastiCache

Vídeos relacionados:

- Deep dive on Amazon EBS
- Optimize your storage performance with Amazon S3
- Modernize apps with purpose-built databases
- Amazon Aurora storage demystified: How it all works
- Amazon DynamoDB deep dive: Advanced design patterns

Exemplos relacionados:

- Driver CSI do Amazon EFS
- Driver CSI do Amazon EBS
- Utilitários do Amazon EFS
- Escalabilidade automática do Amazon EBS
- Exemplos do Amazon S3
- Optimize Data Pattern using Amazon Redshift Data Sharing
- Migrações de bancos de dados
- MS SQL Server - AWS Database Migration Service (AWS DMS) Replication Demo
- Database Modernization Hands On Workshop (Workshop prático de modernização de bancos de dados)
- Amostras da Amazon Neptune

## PERF03-BP02 Avalie as opções de configuração disponíveis para o datastore

Entenda e avalie os vários atributos e opções de configuração disponíveis para seus datastores a fim de otimizar o espaço de armazenamento e o desempenho da workload.

Antipadrões comuns:

- Você só usa um tipo de armazenamento, como o Amazon EBS, para todas as workloads.
- Você usa as IOPS provisionadas para todas as workloads sem testes reais em todos os níveis de armazenamento.
- Você não tem ciência das opções de configuração da solução de gerenciamento de dados escolhida.
- Você conta somente com o aumento do tamanho da instância sem examinar outras opções de configuração.
- Você não testa as características de ajuste de escala do datastore.

Benefícios de estabelecer esta prática recomendada: A exploração e a experimentação das configurações de datastore permitem que você reduza o custo da infraestrutura, melhore a performance e diminua o esforço necessário para manter as workloads.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

Uma workload pode ter um ou mais datastores usados com base nos requisitos de armazenamento e acesso aos dados. Para otimizar a eficiência e o custo do desempenho, você deve avaliar os padrões de acesso aos dados para determinar as configurações apropriadas do datastore. Ao explorar as opções de datastore, leve em consideração vários aspectos, como opções de armazenamento, memória, computação, réplica de leitura, requisitos de consistência, grupo de conexões e opções de armazenamento em cache. Experimente essas várias opções de configuração para melhorar as métricas de eficiência do desempenho.

### Etapas da implementação

- Entenda as configurações atuais (como tipo de instância, tamanho do armazenamento ou versão do mecanismo de banco de dados) do datastore.
- Analise a documentação da AWS e as práticas recomendadas para saber mais sobre as opções de configuração indicadas que podem ajudar a melhorar o desempenho do datastore. As principais opções de datastore a serem consideradas são as seguintes:

| Opção de configuração | Exemplos |
| --- | --- |
| Offloading reads (like read replicas and caching) | • Em tabelas do DynamoDB, é possível descarregar leituras usando o DAX para armazenamento em cache.<br>• Você pode criar um cluster do Amazon ElastiCache for Redis e configurar a aplicação para ler primeiro do cache e voltar para o banco de dados caso o item solicitado não esteja presente.<br>• Todos os bancos de dados relacionais, como Amazon RDS e Aurora, e bancos de dados NoSQL provisionados, como Neptune e Amazon DocumentDB, permitem adicionar réplicas de leitura para descarregar as partes de leitura da workload.<br>• Os bancos de dados de tecnologia sem servidor, como o DynamoDB, ajustarão a escala automaticamente. Verifique se você tem unidades de capacidade de leitura (RCU) suficientes provisionadas para processar a workload. |

| Opção de configuração | Exemplos |
| --- | --- |
| Scaling writes (like partition key sharding or introducing a queue) | • No caso de bancos de dados relacionais, é possível aumentar o tamanho da instância para acomodar uma workload maior, ou aumentar as IOPs provisionadas para permitir um throughput mais alto no armazenamento subjacente.<br>• Também é possível introduzir uma fila na frente do banco de dados, em vez de gravar diretamente nele. Esse padrão permite desacoplar a ingestão do banco de dados e controlar a taxa de fluxo, para que o banco de dados não fique sobrecarregado.<br>• Usar solicitações de gravação em lote em vez de criar muitas transações de curta duração pode ajudar a melhorar o throughput em bancos de dados relacionais de alto volume de gravação.<br>• Os bancos de dados de tecnologia sem servidor, como o DynamoDB, podem ajustar a escala do throughput de gravação automaticamente ou ajustar as unidades da capacidade de gravação (WCU) provisionadas, dependendo do modo da capacidade.<br>• Você ainda pode ter problemas com partições ativas ao atingir os limites de throughput de determinada chave de partição. Isso pode ser mitigado ao escolher uma chave de partição mais igualmente distribuída ou fragmentar a gravação da chave de partição. |

| Opção de configuração | Exemplos |
| --- | --- |
| Policies to manage the lifecycle of your datasets | • Você pode usar o Ciclo de Vida do Amazon S3 para gerenciar os objetos em todo o ciclo de vida. Se os padrões de acesso forem desconhecidos, variáveis ou imprevisíveis, você pode usar o Amazon S3 Intelligent-Tiering, que monitora padrões de acesso e move automaticamente objetos que não foram acessados para níveis de acesso de menor custo. Você pode aproveitar as métricas de Lente de Armazenamento do Amazon S3 para identificar oportunidades de otimização e lacunas no gerenciamento do ciclo de vida.<br>• Gerenciamento do ciclo de vida do Amazon EFS gerencia automaticamente o armazenamento de arquivos para os sistemas de arquivos. |
| Gerenciamento de conexões e agrupamento | • O Amazon RDS Proxy pode ser usado com o Amazon RDS e o Aurora para gerenciar as conexões com o banco de dados.<br>• Bancos de dados de tecnologia sem servidor, como o DynamoDB, não têm conexões associadas a eles, mas considere a capacidade provisionada e as políticas de ajuste de escala automático para lidar com picos na carga. |

- Realize experimentos e testes comparativos em um ambiente que não seja de produção para identificar qual opção de configuração pode atender aos requisitos da workload.
- Depois de experimentar, planeje a migração e valide as métricas de desempenho.
- Use ferramentas de monitoramento da AWS (como o Amazon CloudWatch) e otimização (como a Lente de Armazenamento do Amazon S3) para otimizar continuamente o armazenamento de dados usando o padrão de uso do mundo real.

# Recursos

Documentos relacionados:

- Armazenamento na nuvem com a AWS
- Tipos de volume do Amazon EBS
- Armazenamento do Amazon EC2
- Amazon EFS: performance do Amazon EFS
- Performance do Amazon FSx for Lustre
- Performance do Amazon FSx for Windows File Server
- Documentação do Amazon S3 Glacier: S3 Glacier
- Amazon S3: considerações sobre performance e taxa de solicitação
- Armazenamento na nuvem com a AWS
- Armazenamento na nuvem com a AWS
- Características de E/S do Amazon EBS
- Bancos de dados em nuvem com a AWS
- Armazenamento em cache de banco de dados da AWS
- DynamoDB Accelerator
- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- 10 melhores dicas de desempenho do Amazon Athena
- Práticas recomendadas do Amazon Redshift Spectrum
- Melhores práticas do Amazon DynamoDB

Vídeos relacionados:

- Deep dive on Amazon EBS
- Optimize your storage performance with Amazon S3
- Modernize apps with purpose-built databases
- Amazon Aurora storage demystified: How it all works
- Amazon DynamoDB deep dive: Advanced design patterns

Exemplos relacionados:

- Driver CSI do Amazon EFS
- Driver CSI do Amazon EBS
- Utilitários do Amazon EFS
- Escalabilidade automática do Amazon EBS
- Exemplos do Amazon S3
- Exemplos do Amazon DynamoDB
- AWS Database migration samples
- Database Modernization Workshop (Workshop de modernização de bancos de dados)
- Working with parameters on your Amazon RDS for Postgress DB

## PERF03-BP03 Colete e registre métricas de desempenho do datastore

Acompanhe e registre métricas de desempenho relevantes para o datastore a fim de entender o desempenho de suas soluções de gerenciamento de dados. Essas métricas podem ajudar você a otimizar o datastore, verificar se os requisitos da workload foram atendidos e fornecer uma visão geral clara do desempenho da workload.

Antipadrões comuns:

- Você só usa a pesquisa manual de arquivos de log para métricas.
- Você só publica métricas em ferramentas internas usadas pela equipe e não tem uma imagem abrangente da workload.
- Você só usa as métricas comuns registradas pelo software de monitoramento selecionado.
- Você só revisa as métricas quando há um problema.
- Você só monitora as métricas no sistema e não captura as métricas de uso e acesso aos dados.

Benefícios de estabelecer esta prática recomendada: O estabelecimento de uma linha de base de performance ajuda a compreender o comportamento normal e os requisitos das workloads. Padrões anormais podem ser identificados e depurados mais rapidamente, melhorando a performance e a confiabilidade do datastore.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

Para monitorar a performance dos datastores, você precisa registrar várias métricas de desempenho ao longo de um período. Isso permite detectar anomalias e avaliar o desempenho em relação às métricas de negócios para verificar se as necessidades da workload estão sendo atendidas.

As métricas devem incluir as do sistema subjacente que oferece suporte ao datastore e as do banco de dados. As métricas do sistema subjacente podem incluir métricas de utilização de CPU, memória, armazenamento em disco disponível, E/S de disco, taxa de acertos do cache e entrada e saída da rede, enquanto as métricas do datastore devem incluir transações por segundo, tempos de resposta, uso de índice, bloqueios de tabela, tempos limite de consultas e número de conexões abertas. Esses dados são essenciais para compreender como está a performance da workload e como a solução de gerenciamento de dados é usada. Use essas métricas como parte de uma abordagem orientada por dados para ajustar e otimizar os recursos da workload.

Use ferramentas, bibliotecas e sistemas que registram as medidas de performance relacionadas ao banco de dados.

## Etapas da implementação

1. Identifique as principais métricas de desempenho que o datastore deve monitorar.
    a. Métricas e dimensões do Amazon S3
    b. Métricas de monitoramento para em uma instância do Amazon RDS
    c. Monitorar a carga do banco de dados com o Performance Insights no Amazon RDS
    d. Visão geral do monitoramento aprimorado
    e. Métricas e dimensões do DynamoDB
    f. Monitoramento do DynamoDB Accelerator
    g. Monitoramento do Amazon MemoryDB for Redis com o Amazon CloudWatch
    h. Quais métricas devo monitorar?
    i. Monitoramento da performance do cluster do Amazon Redshift
    j. Métricas e dimensões do Timestream
    k. Métricas do Amazon CloudWatch para Amazon Aurora
    l. Registro em log e monitoramento no Amazon Keyspaces (for Apache Cassandra)
    m. Monitoramento dos recursos do Amazon Neptune

2. Use uma solução aprovada de registro em log e monitoramento para coletar essas métricas. Amazon CloudWatch pode coletar métricas nos recursos na sua arquitetura. Você também pode coletar e publicar métricas personalizadas para descobrir métricas de negócio ou derivadas. Use o CloudWatch ou soluções de terceiros para definir alarmes que indiquem quando os limites são violados.
3. Confira se o monitoramento do datastore pode se beneficiar de uma solução de machine learning que detecta anomalias de performance.
    a. O Amazon DevOps Guru para Amazon RDS fornece visibilidade dos problemas de performance e faz recomendações de ações corretivas.
4. Configure a retenção de dados em sua solução de monitoramento e registro para corresponder às suas metas operacionais e de segurança.
    a. Retenção de dados padrão para métricas do CloudWatch
    b. Retenção de dados padrão para o CloudWatch Logs

## Recursos

Documentos relacionados:

- Armazenamento em cache de banco de dados da AWS
- 10 melhores dicas de desempenho do Amazon Athena
- Práticas recomendadas do Amazon Aurora
- DynamoDB Accelerator
- Melhores práticas do Amazon DynamoDB
- Práticas recomendadas do Amazon Redshift Spectrum
- Performance do Amazon Redshift
- Bancos de dados em nuvem com a AWS
- Insights de performance do Amazon RDS

Vídeos relacionados:

- AWS purpose-built databases
- Amazon Aurora storage demystified: How it all works
- Amazon DynamoDB deep dive: Advanced design patterns

- Best Practices for Monitoring Redis Workloads on Amazon ElastiCache

Exemplos relacionados:

- Level 100: Monitoring with CloudWatch Dashboards
- AWS Dataset Ingestion Metrics Collection Framework
- Amazon RDS Monitoring Workshop

# PERF03-BP04 Implemente estratégias para melhorar o desempenho da consulta no datastore

Implemente estratégias para otimizar os dados e melhorar a consulta de dados a fim de permitir mais escalabilidade e desempenho eficiente para a workload.

Antipadrões comuns:

- Você não particiona dados no datastore.
- Você armazena dados em apenas um formato de arquivo no datastore.
- Você não usa índices no datastore.

Benefícios de estabelecer esta prática recomendada: A otimização da performance dos dados e das consultas ocasiona mais eficiência, menor custo e melhor experiência do usuário.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

A otimização de dados e o ajuste de consultas são aspectos essenciais da eficiência do desempenho em um datastore, pois afetam não só o desempenho como também a capacidade de resposta de toda a workload na nuvem. Consultas não otimizadas podem ocasionar maior uso de recursos e gargalos, o que reduz a eficiência geral de um datastore.

A otimização de dados inclui várias técnicas para garantir o armazenamento e o acesso eficientes aos dados. Esse processo também ajuda a melhorar o desempenho da consulta em um datastore. As principais estratégias incluem particionamento, compactação e desnormalização de dados, que ajudam a otimizá-los para armazenamento e acesso.

### Etapas da implementação

- Entenda e analise as consultas críticas de dados que são realizadas no datastore.
- Identifique as consultas com execução lenta no datastore e use planos de consulta para entender o estado atual delas.
  - Analyzing the query plan in Amazon Redshift
  - Using EXPLAIN and EXPLAIN ANALYZE in Athena
- Implemente estratégias para melhorar o desempenho da consulta. Algumas das principais estratégias incluem:
  - Usar um formato de arquivo colunar (como Parquet ou ORC).
  - Compactar os dados no datastore para reduzir o espaço de armazenamento e a operação de E/S.
  - Particionar os dados para dividi-los em partes menores e reduzir o tempo de verificação dos dados.
    - Partitioning data in Athena
    - Partições e distribuição de dados
  - Indexação de dados nas colunas comuns na consulta.
  - Escolha a operação de junção correta para consulta. Ao unir duas tabelas, especifique a tabela maior no lado esquerdo da junção e a tabela menor no lado direito.
  - Solução de cache distribuído para melhorar a latência e reduzir o número de operações de E/S do banco de dados.
  - Manutenção regular, como execução de estatísticas.
- Experimente e teste estratégias em um ambiente que não seja de produção.

## Recursos

Documentos relacionados:

- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- 10 melhores dicas de desempenho do Amazon Athena
- Armazenamento em cache de banco de dados da AWS
- Melhores práticas para a implementação do Amazon ElastiCache

- Partitioning data in Athena

Vídeos relacionados:

- Optimize Data Pattern using Amazon Redshift Data Sharing
- Optimize Amazon Athena Queries with New Query Analysis Tools

Exemplos relacionados:

- Driver CSI do Amazon EFS

# PERF03-BP05 Implementar padrões de acesso a dados que utilizem cache

Implemente padrões de acesso que possam se beneficiar do armazenamento em cache de dados para recuperação rápida de dados acessados com frequência.

Antipadrões comuns:

- Você armazena em cache dados que mudam com frequência.
- Você depende dos dados em cache como se estivessem armazenados de forma durável e sempre disponíveis.
- Você não leva em conta a consistência dos seus dados em cache.
- Você não monitora a eficiência da sua implementação de cache.

Benefícios de estabelecer esta prática recomendada: armazenar dados em um cache pode melhorar a latência de leitura, throughput de leitura, a experiência do usuário e a eficiência geral, além de reduzir custos.

Nível de risco exposto se esta prática recomendada não for estabelecida: médio

## Orientação para implementação

Um cache é um componente de software ou hardware destinado a armazenar dados para que futuras solicitações dos mesmos dados possam ser atendidas com maior rapidez e eficiência. Os

dados armazenados em um cache podem ser reconstruídos se perdidos, repetindo um cálculo anterior ou obtendo-os de outro armazenamento de dados.

O armazenamento de dados em cache pode ser uma das estratégias mais eficazes para melhorar o desempenho geral da aplicação e reduzir a carga sobre as fontes de dados primárias subjacentes. Os dados podem ser armazenados em cache em vários níveis na aplicação, tais como dentro da aplicação fazendo chamadas remotas, conhecidas como armazenamento em cache do lado do cliente,ou usando um serviço secundário rápido para armazenar os dados, conhecido como armazenamento em cache remoto.

Armazenamento em cache do lado do cliente

Com o armazenamento em cache do lado do cliente, cada cliente (uma aplicação ou serviço que consulta o datastore de back-end) pode armazenar os resultados de suas consultas exclusivas localmente por um período especificado. Isso pode reduzir o número de solicitações na rede para um datastore verificando primeiro o cache do cliente local. Se os resultados não estiverem presentes, a aplicação poderá então consultar o datastore e armazenar esses resultados localmente. Esse padrão permite que cada cliente armazene dados no local mais próximo (o próprio cliente), resultando na menor latência possível. Os clientes também podem continuar a atender algumas consultas quando o datastore de back-end não está disponível, aumentando a disponibilidade geral do sistema.

Uma desvantagem dessa abordagem é que, quando vários clientes estão envolvidos, eles podem armazenar os mesmos dados em cache localmente. Isso resulta no uso de armazenamento duplicado e na inconsistência de dados entre esses clientes. Um cliente pode armazenar em cache os resultados de uma consulta e, um minuto depois, outro cliente pode executar a mesma consulta e obter um resultado diferente.

Armazenamento em cache remoto

Para resolver o problema de dados duplicados entre clientes, um serviço externo rápido, ou cache remoto,pode ser usado para armazenar os dados consultados. Em vez de verificar um datastore local, cada cliente verificará o cache remoto antes de consultar o datastore de back-end. Essa estratégia permite respostas mais consistentes entre clientes, melhor eficiência nos dados armazenados e um volume maior de dados em cache, pois o espaço de armazenamento é dimensionado independentemente dos clientes.

A desvantagem de um cache remoto é que o sistema geral pode ter uma latência maior, pois é necessário um salto de rede adicional para verificar o cache remoto. O cache do lado do cliente pode ser usado junto com o armazenamento em cache remoto para o armazenamento em vários níveis para melhorar a latência.

### Etapas da implementação

1. Identifique bancos de dados, APIs e serviços de rede que poderiam se beneficiar do armazenamento em cache. Serviços que têm workloads de leitura pesadas, uma alta taxa de leitura e gravação ou que são caros para escalar são candidatos ao armazenamento em cache.
   - Armazenamento em cache de banco de dados
   - Ativação do armazenamento em cache da API para melhorar a capacidade de resposta
2. Identifique o tipo apropriado de estratégia de armazenamento em cache que melhor se adapte ao seu padrão de acesso.
   - Estratégias de armazenamento em cache
   - Soluções de armazenamento em cache da AWS
3. Siga Práticas recomendadas de armazenamento em cache para seu armazenamento de dados.
4. Configure uma estratégia de invalidação de cache, como um time-to-live (TTL), para todos os dados que equilibre a atualização dos dados e reduza a pressão sobre o datastore de back-end.
5. Ative recursos como novas tentativas automáticas de conexão, recuo exponencial, tempos limite do lado do cliente e pool de conexões no cliente, se disponíveis, pois eles podem melhorar o desempenho e a confiabilidade.
   - Práticas recomendadas: clientes Redis e Amazon ElastiCache for Redis
6. Monitore a taxa de acertos de cache com uma meta de 80% ou mais. Valores mais baixos podem indicar tamanho insuficiente do cache ou um padrão de acesso que não se beneficia do armazenamento em cache.
   - Quais métricas devo monitorar?
   - Práticas recomendadas para monitorar workloads do Redis no Amazon ElastiCache
   - Monitoramento das práticas recomendadas com Amazon ElastiCache for Redis usando o Amazon CloudWatch
7. Implemente replicação de dados para descarregar as leituras em várias instâncias e melhorar o desempenho e a disponibilidade da leitura de dados.

## Recursos

Documentos relacionados:

- Uso do Amazon ElastiCache Well-Architected Lens

- Monitoramento das práticas recomendadas com Amazon ElastiCache for Redis usando o Amazon CloudWatch
- Quais métricas devo monitorar?
- Whitepaper: Performance at Scale with Amazon ElastiCache (Desempenho em escala com Amazon ElastiCache)
- Desafios e estratégias de armazenamento em cache

Vídeos relacionados:

- Amazon ElastiCache Learning Path (Roteiro de aprendizado do Amazon ElastiCache)
- Design for success with Amazon ElastiCache best practices (Projete para o sucesso com as práticas recomendadas do Amazon ElastiCache)

Exemplos relacionados:

- Como aumentar o desempenho do banco de dados MySQL com Amazon ElastiCache for Redis

# Rede e entrega de conteúdo

A solução de rede ideal para uma workload varia com base nos requisitos de latência, throughput, instabilidade e largura de banda. Restrições físicas, como recursos de usuário ou on-premises, determinam as opções de localização. Essas restrições podem ser compensadas com locais de borda ou posicionamento de recursos.

Na AWS, as redes são virtualizadas e estão disponíveis em vários tipos e configurações diferentes. Desse modo, fica mais fácil atender às suas necessidades de rede. A AWS oferece recursos de produtos (por exemplo, redes avançadas, instâncias otimizadas de rede do Amazon EC2, aceleração de transferências do Amazon S3 e Amazon CloudFront dinâmico) para otimizar o tráfego da rede. A AWS também oferece recursos de rede (por exemplo, roteamento de latência do Amazon Route 53, endpoints da Amazon VPC, AWS Direct Connect e AWS Global Accelerator) para reduzir a distância ou a oscilação da rede.

Essa área de foco compartilha orientações e práticas recomendadas para projetar, configurar e operar soluções eficientes de rede e entrega de conteúdo na nuvem.

Práticas recomendadas

- PERF04-BP01 Compreender como as redes afetam a performance
- PERF04-BP02 Avaliar os recursos de redes disponíveis
- PERF04-BP03 Escolher a conectividade dedicada ou VPN apropriada para a workload
- PERF04-BP04 Usar o balanceamento de carga para distribuir o tráfego em vários recursos
- PERF04-BP05 Escolher os protocolos de rede para melhorar o desempenho
- PERF04-BP06 Escolher o local da workload com base nos requisitos de rede
- PERF04-BP07 Otimizar a configuração da rede com base em métricas

## PERF04-BP01 Compreender como as redes afetam a performance

Analise e entenda como as decisões relacionadas à rede afetam sua workload para fornecer desempenho eficiente e melhor experiência do usuário.

Antipadrões comuns:

- Todo o tráfego flui por meio dos datacenters existentes.

- Você direciona todo o tráfego por meio de firewalls centrais em vez de usar ferramentas de segurança de rede nativas da nuvem.
- Você provisiona conexões do AWS Direct Connect sem entender os requisitos reais de uso.
- Você não considera as características da workload e a sobrecarga da criptografia ao definir suas soluções de redes.
- Você usa conceitos e estratégias de on-premises para soluções de redes na nuvem.

Benefícios de estabelecer esta prática recomendada: a compreensão de como as redes afetam a performance da workload ajuda a identificar gargalos potenciais, a melhorar a experiência dos usuários, a aumentar a confiabilidade e a reduzir a manutenção operacional à medida que a workload muda.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

A rede é responsável pela conectividade entre os componentes da aplicação, os serviços de nuvem, as redes de borda e os dados on-premises, portanto ela pode afetar significativamente a performance da workload. Além da performance da workload, a experiência dos usuários também é afetada pela latência da rede, a largura de banda, os protocolos, a localização, a congestão da rede, a variação de latência (jitter), o throughput e as regras de roteamento.

Ter uma lista documentada dos requisitos de rede da workload, incluindo latência, tamanho de pacotes, regras de roteamento, protocolos e padrões de tráfego compatíveis. Analise as soluções de redes disponíveis e identifique os serviços que atendem às características de redes da sua workload. É possível recriar as redes baseadas na nuvem rapidamente, portanto, é necessário evoluir sua arquitetura de rede ao longo do tempo para melhorar a eficiência da performance.

### Etapas da implementação:

1. Defina e documente os requisitos de desempenho da rede, incluindo métricas como latência da rede, largura de banda, protocolos, locais, padrões de tráfego (picos e frequência), throughput, criptografia, inspeção e regras de roteamento.
2. Saiba mais sobre os principais serviços de rede da AWS, como VPCs, O AWS Direct Connect, Elastic Load Balancing (ELB) e Amazon Route 53.
3. Capture as seguintes características principais de rede:

| Características | Ferramentas e métricas |
| --- | --- |
| Características básicas de rede | • Logs de fluxo da VPC<br>• Logs de fluxo do AWS Transit Gateway<br>• Métricas do AWS Transit Gateway<br>• Métricas do AWS PrivateLink |
| Características da rede de aplicações | • Elastic Fabric Adapter<br>• Métricas do AWS App Mesh<br>• Métricas do Amazon API Gateway |
| Características da rede de borda | • Métricas do Amazon CloudFront<br>• Métricas do Amazon Route 53<br>• Métricas do AWS Global Accelerator |
| Características da rede híbrida | • Métricas do AWS Direct Connect<br>• Métricas do AWS Site-to-Site VPN<br>• Métricas do AWS Client VPN<br>• Métricas da WAN da Nuvem AWS |
| Características da rede de segurança | • Métricas do AWS Shield,AWS WAF e AWS Network Firewall |
| Características de rastreamento | • AWS X-Ray<br>• VPC Reachability Analyzer<br>• Network Access Analyzer<br>• Amazon Inspector<br>• Amazon CloudWatch RUM |

4. Realize o teste comparativo e de performance da rede:
   a. Realize o teste comparativo do throughput da rede, pois alguns fatores podem afetar o desempenho da rede do Amazon EC2 quando as instâncias estão na mesma VPC. Meça a largura de banda da rede entre as instâncias do Amazon EC2 Linux na mesma VPC.
   b. Execute testes de carga para experimentar soluções e opções de redes.

# Recursos

Documentos relacionados:

- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Adaptador de Rede Elástica (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Transit Gateway
- Fazer a transição para o encaminhamento por latência no Amazon Route 53
- Endpoints da VPC
- Logs de fluxo da VPC

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (Conectividade com a AWS e arquiteturas de rede híbrida da AWS)
- Optimizing Network Performance for Amazon EC2 Instances (Otimização da performance da rede para instâncias do Amazon EC2)
- Improve Global Network Performance for Applications (Melhorar a performance da rede global para aplicações)
- EC2 Instances and Performance Optimization Best Practices (Práticas recomendadas para instâncias do EC2 e otimização da performance)
- Optimizing Network Performance for Amazon EC2 Instances (Otimização da performance da rede para instâncias do Amazon EC2)
- Networking best practices and tips with the Well-Architected Framework (Práticas recomendadas e dicas de redes com o Well-Architected Framework)
- AWS networking best practices in large-scale migrations (Práticas recomendadas da AWS em migrações de grande escala)

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS

# PERF04-BP02 Avaliar os recursos de redes disponíveis

Avalie recursos de rede na nuvem que possam melhorar o desempenho. Meça o impacto desses recursos por meio de testes, métricas e análises. Por exemplo, aproveite os recursos de rede que estão disponíveis para reduzir a latência, a distância ou a instabilidade da rede.

Antipadrões comuns:

- Você permanece em uma Região, pois é onde sua sede está fisicamente localizada.
- Você usa firewalls em vez de grupos de segurança para filtrar o tráfego.
- Você quebra o TLS para inspeção de tráfego em vez de confiar em grupos de segurança, políticas de endpoint e outras funcionalidades nativas da nuvem.
- Você só usa segmentação baseada em sub-rede em vez de grupos de segurança.

Benefícios de estabelecer esta prática recomendada: avaliar todos os recursos e opções de serviços pode aumentar a performance da workload, reduzir o custo da infraestrutura, diminuir o esforço necessário para manter sua workload e aumentar sua postura geral de segurança. É possível utilizar a espinha dorsal da AWS para garantir a experiência ideal de redes para os clientes.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

A AWS oferece serviços como AWS Global Accelerator e o Amazon CloudFront que podem ajudar a melhorar o desempenho da rede, enquanto a maioria dos serviços da AWS tem recursos de produto (como a Aceleração de Transferências do Amazon S3) para otimizar o tráfego de rede.

Analise quais opções de configuração de rede estão disponíveis e como elas poderiam afetar a workload. A otimização do desempenho depende da compreensão de como essas opções interagem com sua arquitetura e do impacto que elas terão no desempenho medido e na experiência do usuário.

## Etapas da implementação

- Crie uma lista de componentes da workload.
  - Considere o uso de Nuvem AWS WAN para criar, gerenciar e monitorar a rede da sua organização ao criar uma rede global unificada.
  - Monitore suas redes globais e centrais com métricas do Amazon CloudWatch Logs. Utilize o Amazon CloudWatch RUM, que fornece insights para ajudar a identificar, entender e aprimorar a experiência digital dos usuários.
  - Visualize a latência agregada da rede entre Regiões da AWS e Zonas de Disponibilidade, bem como dentro de cada Zona de Disponibilidade, usando AWS Network Manager para obter informações sobre como o desempenho da sua aplicação se relaciona com o desempenho da rede da AWS subjacente.
  - Use uma ferramenta de banco de dados de gerenciamento de configurações (CMDB) existente ou uma ferramenta como o AWS Config para criar um inventário de sua workload e como ela é configurada.
- Se for uma workload existente, identifique e documente a referência para suas métricas de performance, focando nos gargalos e nas áreas de melhoria. As métricas de rede associadas a performance vão variar de acordo com a workload com base nos requisitos comerciais e nas características da workload. Como ponto de partida, a análise dessas métricas pode ser importante para sua workload: largura de banda, latência, perda de pacotes, instabilidade da rede e retransmissões.
- Se a workload for nova, realize testes de carga para identificar gargalos de performance.
- Para os gargalos de performance que identificar, analise as opções de configuração para suas soluções a fim de identificar oportunidades de melhoria da performance. Confira as seguintes principais opções e recursos de rede:

| Oportunidade de melhoria | Solução |
| --- | --- |
| Caminho ou rotas de rede | Use o Network Access Analyzer para identific ar caminhos ou rotas. |
| Protocolos de rede | Consulte PERF04-BP05 Escolher os protocolo s de rede para melhorar o desempenho |
| Topologia de rede | Avalie suas concessões de operação e performance entre Emparelhamento de VPC |

| Oportunidade de melhoria | Solução |
| --- | --- |
| | e AWS Transit Gateway ao conectar várias contas. O AWS Transit Gateway simplifica a forma como você interconecta todas as suas VPCs, que podem se estender por milhares de Contas da AWS e até redes on-premises. Compartilhe seu AWS Transit Gateway entre várias contas usando o AWS Resource Access Manager.<br><br>Consulte PERF04-BP03 Escolher a conectividade dedicada ou VPN apropriada para a workload |

| Oportunidade de melhoria | Solução |
| --- | --- |
| Serviços de rede | O AWS Global Accelerator é um serviço de rede que melhora a performance do tráfego dos usuários em até 60% usando a infraestr utura de rede global da AWS.<br><br>O Amazon CloudFront pode melhorar o desempenho da entrega e da latência de conteúdo da workload globalmente.<br><br>Use o Lambda@edge para executar funções que personalizam o conteúdo que o CloudFront entrega mais perto dos usuários, reduzem a latência e melhoram o desempenh o.<br><br>O Amazon Route 53 oferece opções de roteamento baseado em latência, roteamento por geolocalização, roteamento por geoproxim idadee aos roteamento baseado em IP para ajudar a melhorar a performance da workload para um público global. Identifique qual opção de roteamento otimizaria o desempenho da workload analisando o tráfego dela e a localização do usuário quando ela for distribuí da globalmente. |

| Oportunidade de melhoria | Solução |
| --- | --- |
| Recursos do atributo de armazenamento | Aceleração de Transferências do Amazon S3) é um recurso que permite que usuários externos se beneficiem de otimizações de rede do CloudFront a fim de fazer upload de dados no Amazon S3. Isso melhora a capacidade de transferir grandes quantidades de dados com origem em locais remotos que não têm conectividade dedicada com a Nuvem AWS.<br><br>Pontos de acesso multirregionais no Amazon S3 replicam conteúdo para várias regiões e simplificam a workload ao proporcionar um ponto de acesso. Quando um ponto de acesso multirregional é usado, você pode solicitar ou gravar dados no Amazon S3 com o serviço identificando o bucket de menor latência. |

| Oportunidade de melhoria | Solução |
| --- | --- |
| Atributos de recursos computacionais | Interfaces de rede elástica (ENA) usadas por instâncias do Amazon EC2, contêineres e funções do Lambda são limitadas por fluxo. Revise seus grupos de posicionamento para otimizar o throughput de rede do EC2. Para evitar gargalos em uma abordagem por fluxo, projete sua aplicação para usar vários fluxos. Para monitorar e obter visibilidade de suas métricas de rede relacionadas à computaçã o, use o CloudWatch Metrics e a ethtool. O comando `ethtool` está incluído no driver da ENA e expõe métricas adicionais relacionadas à rede que podem ser publicadas como uma métrica personalizada no CloudWatch.<br><br>Adaptadores de rede elástica (ENA) da Amazon proporcionam ainda mais otimizaçã o ao oferecer mais throughput para suas instâncias em um grupo com posicionamento em cluster.<br><br>Elastic Fabric Adapter (EFA) é uma interface de rede para instâncias do Amazon EC2 que permite executar workloads que exigem altos níveis de comunicação entre nós em grande escala na AWS.<br><br>Instâncias otimizadas para Amazon EBS usam uma pilha de configuração otimizada e fornecem capacidade adicional e dedicada para aumentar a E/S do Amazon EBS. |

## Recursos

Documentos relacionados:

- Instâncias otimizadas para Amazon EBS
- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Adaptador de Rede Elástica (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- AWS Transit Gateway
- Fazer a transição para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints
- Logs de fluxo da VPC

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (Conectividade com a AWS e arquiteturas de rede híbrida da AWS)
- Optimizing Network Performance for Amazon EC2 Instances (Otimização da performance da rede para instâncias do Amazon EC2)
- AWS Global Accelerator

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS

## PERF04-BP03 Escolher a conectividade dedicada ou VPN apropriada para a workload

Quando a conectividade híbrida é necessária para conectar recursos on-premises e na nuvem, provisione a largura de banda adequada para atender aos requisitos de performance. Estime os requisitos de largura de banda e de latência para a workload híbrida. Esses números determinarão seus requisitos de dimensionamento.

Antipadrões comuns:

- Você só avalia as soluções de VPN para seus requisitos de criptografia de rede.
- Você não avalia as opções de backup ou de conectividade redundante.
- Você não identifica todos os requisitos da workload (necessidades de criptografia, protocolo, largura de banda e tráfego).

Benefícios de estabelecer esta prática recomendada: Selecionar e configurar soluções de conectividade apropriadas aumentará a confiabilidade da workload e maximizará a performance. A identificação dos requisitos da workload, o planejamento antecipado e a avaliação das soluções híbridas podem minimizar alterações dispendiosas da rede física e despesas operacionais, e aumentará seu time-to-value.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

Desenvolva uma arquitetura de rede híbrida com base em seus requisitos de largura de banda. O AWS Direct Connect permite que você conecte sua rede on-premises de forma privada com a AWS. Isso lhe dará segurança quando você precisar de largura de banda alta e baixa latência com uma performance consistente. Uma conexão VPN estabelece uma conexão segura pela internet. Ela é usada quando apenas uma conexão temporária é necessária, quando o custo é um fator ou como uma contingência enquanto se espera que uma conectividade de rede física resiliente seja estabelecida durante o uso do AWS Direct Connect.

Se seus requisitos de largura de banda forem altos, considere vários serviços do AWS Direct Connect ou de VPN. O tráfego pode ser balanceado entre os serviços, embora não recomendamos o balanceamento de carga entre o AWS Direct Connect e a VPN devido às diferenças de latência e largura de banda.

### Etapas da implementação

1. Calcule os requisitos de largura de banda e latência de suas aplicações existentes.
   a. Para workloads existentes que estão sendo migradas para a AWS, utilize os dados de seus sistemas de monitoramento de rede internos.
   b. Para workloads novas ou existentes para as quais não há dados de monitoramento, consulte os proprietários do produto para determinar métricas de performance adequadas e fornecer uma experiência do usuário satisfatória.

2. Escolha uma conexão dedicada ou VPN como sua opção de conectividade. Com base em todos os requisitos da workload (necessidades de criptografia, largura de banda e tráfego), é possível escolher o AWS Direct Connect ou a AWS VPN (ou ambos). O diagrama a seguir ajudará você a escolher o tipo de conexão apropriada.
    a. O AWS Direct Connect fornece conectividade dedicada ao ambiente da AWS, de 50 Mbps a 100 Gbps, usando conexões dedicadas ou conexões hospedadas. Isso permite que você tenha latência gerenciada e controlada, além de largura de banda provisionada para que a workload possa se conectar de forma eficiente com outros ambientes. Com os parceiros do AWS Direct Connect, é possível ter conectividade completa para vários ambientes, fornecendo uma rede estendida com performance consistente. A AWS oferece escalabilidade da largura de banda da conexão direta usando o grupo de agregação nativo (LAG) de 100 Gbps ou o BGP equal-cost multipath (ECMP).
    b. A AWS Site-to-Site VPN fornece um serviço de VPN gerenciada compatível com o protocolo de segurança da internet (IPsec). Quando uma conexão VPN é criada, cada conexão VPN inclui dois túneis para alta disponibilidade.
3. Siga a documentação da AWS para escolher uma opção de conectividade apropriada:
    a. Se você decidir usar o AWS Direct Connect, selecione a largura de banda apropriada para sua conectividade.
    b. Se você usar uma AWS Site-to-Site VPN em vários locais para se conectar a uma Região da AWS, use uma conexão de Site-to-Site VPN acelerada para melhorar a performance da rede.
    c. Se o design da sua rede consistir em uma conexão VPN IPSec no AWS Direct Connect, considere o uso de VPN de IP privado para melhorar a segurança e conseguir segmentação. A AWS Site-to-Site Private IP VPN é implantada sobre a interface virtual de trânsito (VIF).
    d. O AWS Direct Connect SiteLink permite criar conexões redundantes e de baixa latência entre seus datacenters em todo o mundo, enviando dados pelo caminho mais rápido entre os locais do AWS Direct Connect, contornando Regiões da AWS.
4. Valide sua configuração de conectividade antes de implantá-la na produção. Execute testes de segurança e performance para garantir que ela atenda aos requisitos de largura de banda, confiabilidade, latência e conformidade.
5. Monitore regularmente a performance e o uso da conectividade e otimize, se necessário.

Fluxograma de desempenho determinístico

# Recursos

Documentos relacionados:

- Network Load Balancer
- Produtos de redes com a AWS
- AWS Transit Gateway
- Fazer a transição para o encaminhamento por latência no Amazon Route 53
- Endpoints da VPC
- Site-to-Site VPN
- Building a Scalable and Secure Multi-VPC AWS Network Infrastructure (Criação de uma infraestrutura de rede da AWS de várias VPCs escaláveis e seguras)
- AWS Direct Connect
- Client VPN

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (Conectividade com a AWS e arquiteturas de rede híbrida da AWS)
- Optimizing Network Performance for Amazon EC2 Instances (Otimização da performance da rede para instâncias do Amazon EC2)
- AWS Global Accelerator
- AWS Direct Connect
- AWS Transit Gateway Connect
- Soluções de VPN
- Segurança com as soluções de VPN

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS

# PERF04-BP04 Usar o balanceamento de carga para distribuir o tráfego em vários recursos

Distribua o tráfego entre vários recursos e serviços para permitir que sua workload aproveite a elasticidade que a nuvem oferece. Também é possível usar o balanceamento de carga para descarregar a terminação de criptografia a fim de melhorar a performance, a confiabilidade e gerenciar e rotear o tráfego de maneira eficaz.

Antipadrões comuns:

- Você não considera os requisitos da workload ao escolher o tipo de balanceador de carga.
- Você não utiliza os recursos do balanceador de carga para otimização do desempenho.
- A workload é exposta diretamente à internet sem um balanceador de carga.
- Você roteia todo o tráfego da Internet por meio de balanceadores de carga existentes.
- Você usa o balanceamento de carga TCP genérico e faz com que cada nó de computação lide com a criptografia SSL.

Benefícios de estabelecer esta prática recomendada: Um balanceador de carga lida com a carga variável do tráfego da sua aplicação em uma única Zona de Disponibilidade ou em várias Zonas de Disponibilidade e permite alta disponibilidade, ajuste de escala automático e melhor utilização de sua workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

Os balanceadores de carga atuam como o ponto de entrada para sua workload, a partir do qual distribuem o tráfego para seus destinos de back-end, como instâncias de computação ou contêineres, para melhorar a utilização.

Escolher o tipo certo de balanceador de carga é a primeira etapa para otimizar sua arquitetura. Comece listando as características da workload, como protocolo (como TCP, HTTP, TLS ou WebSockets), o tipo de destino (como instâncias, contêineres ou tecnologia sem servidor), requisitos da aplicação (como conexões de execução longa, autenticação de usuários ou adesão) e posicionamento (como região, zona local, Outpost ou isolamento por zona).

A AWS fornece vários modelos para que suas aplicações usem o balanceamento de carga. O Application Load Balancer é o mais adequado para balanceamento de carga de tráfego HTTP e

HTTPS, e oferece roteamento avançado de solicitação direcionado para a entrega de arquiteturas de aplicações modernas, inclusive microsserviços e contêineres.

O Network Load Balancer é o mais adequado para o balanceamento de carga de tráfego TCP que exija performance extrema. Ele é capaz de processar milhões de solicitações por segundo enquanto mantém latências ultrabaixas, e também é otimizado para lidar com padrões de tráfego súbitos e voláteis.

O Elastic Load Balancing oferece gerenciamento integrado de certificados e descriptografia SSL/TLS, o que proporciona a flexibilidade de gerenciar centralmente as configurações SSL do balanceador de carga e descarregar de sua workload as interações com uso intenso de CPU.

Depois de escolher o balanceador de carga certo, você pode começar a utilizar seus recursos para reduzir a quantidade de esforço que seu back-end precisa fazer para atender o tráfego.

Por exemplo, ao usar tanto o Application Load Balancer (ALB) como o Network Load Balancer (NLB), é possível realizar o descarregamento de criptografia SSL/TLS, que é uma oportunidade de evitar que o handshake TLS com uso intenso da CPU seja concluído pelos destinos e também melhorar o gerenciamento de certificados.

Ao configurar o descarregamento de SSL/TLS no balanceador de carga, ele se torna responsável pela criptografia do tráfego de e para os clientes enquanto entrega o tráfego não criptografado aos back-ends, liberando os recursos de back-end e melhorando o tempo de resposta para os clientes.

O Application Load Balancer também pode fornecer tráfego HTTP/2 sem precisar comportá-lo em seus destinos. Essa simples decisão pode melhorar o tempo de resposta da aplicação, já que o HTTP/2 usa conexões TCP de forma mais eficiente.

Os requisitos de latência da workload devem ser considerados ao definir a arquitetura. Como exemplo, se você tiver uma aplicação sensível à latência, poderá decidir usar o Network Load Balancer, que oferece latências extremamente baixas. Como alternativa, você pode decidir aproximar a workload dos clientes utilizando o Application Load Balancer em zonas locais da AWS ou mesmo o AWS Outposts.

Outra consideração para workloads sensíveis à latência é o balanceamento de carga entre zonas. Com o balanceamento de carga entre zonas, cada nó do balanceador de carga distribui o tráfego entre os destinos registrados em todas as Zonas de Disponibilidade habilitadas.

Use o Auto Scaling integrado ao balanceador de carga. Um dos principais aspectos de um sistema com desempenho eficiente está relacionado ao dimensionamento correto dos recursos de back-

end. Para fazer isso, é possível utilizar as integrações do balanceador de carga para os recursos de destino de back-end. Ao usar a integração do balanceador de carga com os grupos do Auto Scaling, os destinos serão adicionados ou removidos do balanceador de carga conforme exigido em resposta ao tráfego recebido. Os balanceadores de carga também podem se integrar com o Amazon ECS e o Amazon EKS para workloads em contêineres.

- Amazon ECS: balanceamento de carga do serviço
- Balanceamento de carga da aplicação no Amazon EKS
- Balanceamento de carga da rede no Amazon EKS

## Etapas da implementação

- Defina seus requisitos de balanceamento de carga, incluindo excelente volume, disponibilidade e escalabilidade de aplicações.
- Escolha o tipo certo de balanceador de carga para sua aplicação.
  - Use o Application Load Balancer para workloads HTTP/HTTPS.
  - Use o Network Load Balancer para workloads que não são HTTP que executam TCP ou UDP.
  - Use uma combinação de ambos (ALB como alvo do NLB) se você quiser aproveitar os recursos de ambos os produtos. Por exemplo, é possível fazer isso se você quiser usar os IPs estáticos do NLB junto com o roteamento baseado em cabeçalho HTTP do ALB, ou se quiser expor a workload HTTP em um AWS PrivateLink.
  - Para uma comparação completa dos balanceadores de carga, consulte Comparação de produtos do ELB.
- Use o descarregamento de SSL/TLS, se possível.
  - Configure receptores HTTPS/TLS com o Application Load Balancer e o Network Load Balancer integrados com o AWS Certificate Manager.
  - Observe que algumas workloads podem exigir criptografia completa por motivos de conformidade. Nesse caso, é um requisito para permitir a criptografia nos destinos.
  - Para práticas recomendadas de segurança, consulte SEC09-BP02 Aplicar a criptografia em trânsito.
- Escolha o algoritmo de roteamento certo (apenas ALB).
  - O algoritmo de roteamento pode fazer a diferença em como os destinos de back-end são bem-utilizados e, portanto, na forma como afetam o desempenho. Por exemplo, a ALB fornece duas opções para algoritmos de roteamento:

  - Solicitações menos urgentes: use para obter uma melhor distribuição de carga para seus destinos de back-end em casos nos quais as solicitações para a aplicação variam em complexidade ou os destinos variam na capacidade de processamento.
  - Round robin: use quando as solicitações e os destinos forem semelhantes, ou se você precisar distribuir as solicitações igualmente entre os destinos.
- Considere isolamento por zona ou entre zonas.
  - Desative a opção entre zonas (isolamento por zona) para melhorias de latência e domínios com falha de zona. Ele está desativado por padrão no NLB e no ALB. Você pode desativá-lo por grupo-alvo.
  - Ative a opção entre zonas para maior disponibilidade e flexibilidade. Por padrão, a opção entre zonas está ativada para o ALB. No NLB, você pode ativá-la por grupo-alvo..
- Ative as manutenções de funcionamento de HTTP para as workloads HTTP (apenas ALB). Com esse recurso, o balanceador de carga pode reutilizar as conexões de back-end até expirar o tempo limite da manutenção de funcionamento, melhorando a solicitação HTTP e o tempo de resposta, além de reduzir a utilização de recursos nos destinos de back-end. Para obter detalhes sobre como fazer isso para Apache e Nginx, consulte Quais são as configurações ideais para usar o Apache ou o NGINX como servidor de back-end para o ELB?
- Ative o monitoramento do balanceador de carga.
  - Ative os logs de acesso para o Application Load Balancer e o Network Load Balancer.
  - Os principais campos a considerar para o ALB são `request_processing_time`, o `request_processing_timee` o `response_processing_time`.
  - Os principais campos a considerar para o NLB são `connection_time` e o `tls_handshake_time`.
  - Esteja pronto para consultar os logs quando precisar deles. Você pode usar o Amazon Athena para consultar tanto os logs do ALB e os logs do NLB.
  - Crie alarmes para métricas relacionadas ao desempenho, como `TargetResponseTime` para o ALB.

## Recursos

Documentos relacionados:

- Comparação de produtos do ELB
- AWS Global Infrastructure (Infraestrutura global da AWS)

- Improving Performance and Reducing Cost Using Availability Zone Affinity (Melhorar o desempenho e reduzir os custos usando a afinidade de zona de disponibilidade)
- Step by step for Log Analysis with Amazon Athena (Passo a passo para a análise de logs com o Amazon Athena)
- Querying Application Load Balancer logs (Consulta de logs do Application Load Balancer)
- Monitor your Application Load Balancers (Monitore o Application Load Balancers)
- Monitor your Network Load Balancer (Monitore o Network Load Balancer)
- Use Elastic Load Balancing to distribute traffic across the instances in your Auto Scaling group (Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling)

Vídeos relacionados:

- AWS re:Invent 2018: Elastic Load Balancing: Deep Dive and Best Practices (AWS re:Invent 2018: Elastic Load Balancing: aprofundamento e práticas recomendadas)
- AWS re:Invent 2021 - How to choose the right load balancer for your AWS workloads (AWS re:Invent 2021: como escolher o balanceador de carga certo para suas workloads da AWS)
- AWS re:Inforce 2022 - How to use Elastic Load Balancing to enhance your security posture at scale (AWS re:Inforce 2022: como usar o Elastic Load Balancing para melhorar seu procedimento de segurança em escala)
- AWS re:Invent 2019: Get the most from Elastic Load Balancing for different workloads (AWS re:Invent 2019: aproveite ao máximo o Elastic Load Balancing para diferentes workloads)

Exemplos relacionados:

- CDK and AWS CloudFormation samples for Log Analysis with Amazon Athena (Exemplos de CDK e AWS CloudFormation para análise de log com o Amazon Athena)

## PERF04-BP05 Escolher os protocolos de rede para melhorar o desempenho

Tome decisões sobre protocolos de comunicação entre sistemas e redes com base no impacto na performance da workload.

Há uma relação entre latência e largura de banda para alcançar o throughput. Por exemplo, se a transferência de arquivos estiver usando TCP (Protocolo de Controle de Transmissão), latências mais altas provavelmente reduzirão o throughput geral. Existem abordagens para corrigir isso com ajuste de TCP e protocolos de transferência otimizados, mas uma solução é usar o User Datagram Protocol (UDP, protocolo de datagrama de usuário).

Antipadrões comuns:

- Você usa TCP para todas as workloads, independentemente dos requisitos de performance.

Benefícios de estabelecer esta prática recomendada: verificar se um protocolo apropriado é usado para comunicação entre usuários e componentes da workload ajuda a melhorar a experiência geral do usuário para as aplicações. Por exemplo, o UDP sem conexão permite alta velocidade, mas não oferece retransmissão ou alta confiabilidade. TCP é um protocolo completo, mas requer maior sobrecarga para processar os pacotes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

Se você puder escolher protocolos diferentes para sua aplicação e tiver experiência nessa área, otimize sua aplicação e a experiência do usuário final usando um protocolo diferente. Observe que essa abordagem apresenta dificuldades significativas e só deve ser experimentada se você tiver otimizado sua aplicação de outras maneiras primeiro.

Uma consideração primária para melhorar o desempenho da workload é entender os requisitos de latência e throughput e escolher os protocolos de rede que otimizam o desempenho.

Quando considerar o uso do TCP

O TCP oferece entrega de dados confiável e pode ser usado para comunicação entre componentes da workload em que a confiabilidade e a entrega garantida de dados é importante. Muitas aplicações baseadas na web dependem de protocolos baseados em TCP, como HTTP e HTTPS, para abrir soquetes TCP para comunicação entre componentes da aplicação. A transferência de dados por e-mail e arquivo são aplicações comuns que também usam o TCP, pois é um mecanismo de transferência simples e confiável entre os componentes da aplicação. Usar o TLS com TCP pode adicionar sobrecarga à comunicação, o que pode resultar em maior latência e redução de throughput, mas traz a vantagem da segurança. A sobrecarga vem principalmente da sobrecarga adicionada do processo de handshake, que pode levar várias idas e voltas para ser concluído.

Quando o handshake for concluído, a sobrecarga da criptografia e descriptografia de dados será relativamente pequena.

Quando considerar o uso do UDP

O UDP é um protocolo sem conexão e, portanto, é adequado para aplicações que precisam de uma transmissão rápida e eficiente, como log, monitoramento e dados de VoIP. Além disso, considere usar o UDP se você tiver componentes da workload que respondam a pequenas consultas de grandes números de clientes para garantir um desempenho ideal da workload. O Datagram Transport Layer Security (DTLS) é o equivalente UDP do Transport Layer Security (TLS). Ao usar DTLS com UDP, a sobrecarga vem da criptografia e descriptografia de dados, já que o processo de handshake é simplificado. O DTLS também adiciona uma pequena quantidade de sobrecarga aos pacotes de UDP, já que inclui campos adicionais para indicar os parâmetros de segurança e detectar violações.

Quando considerar o uso do SRD

O SRD (datagrama confiável escalável) é um protocolo de transporte de rede otimizado para workloads de alto throughput devido à sua capacidade de fazer o balanceamento de carga do tráfego em vários caminhos e de se recuperar rapidamente de quedas de pacote ou falhas no link. Assim, o SRD é melhor nos casos de workloads de computação de alta performance (HPC) que exigem comunicação de alto throughput e baixa latência entre os nós de computação. Isso pode incluir tarefas de processamento paralelas, como simulação, modelagem e análise de dados que envolvem uma grande quantidade de transferência de dados entre os nós.

## Etapas da implementação

1. Use o AWS Global Accelerator e o AWS Transfer Family para melhorar o throughput de suas aplicações de transferência de arquivos online. O serviço AWS Global Accelerator ajuda você a obter baixa latência entre os dispositivos cliente e a workload na AWS. Com o AWS Transfer Family, é possível usar protocolos baseados em TCP, como SFTP (Protocolo de transferência de arquivos de Secure Shell) e FTPS (Protocolo de transferência de arquivos por SSL), para escalar e gerenciar com segurança as transferências de arquivos para os serviços de armazenamento da AWS.
2. Use a latência de rede para determinar se o TCP é adequado para comunicação entre os componentes da workload. Se a latência de rede entre a aplicação cliente e o servidor for alta, o handshake de três vias do TCP pode levar um tempo, afetando, assim, a capacidade de resposta da aplicação. Métricas como tempo até o primeiro byte (TTFB) e tempo de ida e volta (RTT)

podem ser usadas para medir a latência da rede. Se sua workload fornece conteúdo dinâmico aos usuários, considere usar o Amazon CloudFront, que estabelece uma conexão persistente com cada origem de conteúdo dinâmico para remover o tempo de configuração da conexão que, de outra forma, diminuiria a velocidade de cada solicitação do cliente.

3. Usar TLS com TCP ou UDP pode resultar em maior latência e menor throughput para a workload devido ao impacto da criptografia e descriptografia. Para essas workloads, considere o descarregamento de SSL/TLS no Elastic Load Balancing para melhorar o desempenho da workload, permitindo que o balanceador de carga lide com o processo de criptografia e descriptografia de SSL/TLS em vez de deixar que as instâncias de back-end façam isso. Isso pode ajudar a reduzir a utilização da CPU nas instâncias de back-end, o que pode melhorar o desempenho e aumentar a capacidade.
4. Use o Network Load Balancer (NLB) para implantar serviços que dependem do protocolo UDP, como autenticação e autorização, registro em log, DNS, IoT e mídia de streaming, visando melhorar o desempenho e a confiabilidade da workload. O NLB distribui o tráfego de UDP de entrada em vários destinos, permitindo escalar a workload horizontalmente, aumentar a capacidade e reduzir a sobrecarga de um único destino.
5. Para suas workloads de computação de alta performance (HPC), considere usar a funcionalidade do Adaptador de Rede Elástica (ENA) Express, que usa o protocolo SRD para melhorar o desempenho da rede fornecendo uma maior largura de banda de fluxo único (25 Gbps) e menor latência final (99,9 percentil) para o tráfego de rede entre instâncias do EC2.
6. Use o Application Load Balancer(ALB) para rotear e balancear a carga do tráfego de gRPC (Chamadas de procedimento remoto) entre os componentes da workload ou entre os serviços e clientes com gRPC habilitadas. As gRPC usam o protocolo HTTP/2 baseado em TCP para transporte e oferece benefícios de desempenho, como pegada de rede mais leve, compactação, serialização binária eficiente, suporte para várias linguagens e streaming bidirecional.

## Recursos

Documentos relacionados:

- Instâncias otimizadas para Amazon EBS
- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2

- Como habilitar a rede avançada com o Adaptador de Rede Elástica (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- AWS Transit Gateway
- Fazer a transição para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints
- Logs de fluxo da VPC

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (Conectividade com a AWS e arquiteturas de rede híbrida da AWS)
- Optimizing Network Performance for Amazon EC2 Instances (Otimização da performance da rede para instâncias do Amazon EC2)

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS

## PERF04-BP06 Escolher o local da workload com base nos requisitos de rede

Avalie as opções para o posicionamento de recursos visando reduzir a latência da rede e melhorar o throughput, proporcionando uma ótima experiência do usuário ao reduzir os tempos de carregamento da página e de transferência de dados.

Antipadrões comuns:

- Você consolida todos os recursos da workload em uma única localização geográfica.
- Você escolhe a Região mais próxima ao seu local, mas não ao usuário final da workload.

Benefícios de estabelecer esta prática recomendada: A experiência do usuário é muito afetada pela latência entre o usuário e sua aplicação. Ao usar Regiões da AWS adequadas e a rede global

privada da AWS, você pode reduzir a latência e oferecer uma melhor experiência aos usuários remotos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

Recursos, como instâncias do Amazon EC2, são colocados em zonas de disponibilidade em Regiões da AWS, zonas locais da AWS, AWS Outpostsou AWS Wavelength . A escolha desse local influencia o throughput e a latência da rede de determinado local do usuário. Serviços de borda, como Amazon CloudFront e o AWS Global Accelerator também podem ser usados para melhorar o desempenho da rede, seja armazenando o conteúdo em cache nos locais da borda ou oferecendo aos usuários um ótimo caminho para a workload por meio da rede global da AWS.

O Amazon EC2 oferece grupos de posicionamento para redes. Um grupo de posicionamento é um agrupamento lógico de instâncias para diminuir a latência. O uso de grupos de posicionamento com tipos de instância compatíveis e um Adaptador de Rede Elástica (ENA) permite que as workloads participem de uma rede de baixa latência, com oscilação reduzida e de 25 Gbps. Recomenda-se o uso de grupos de posicionamento para workloads que se beneficiam de baixa latência de rede, alto throughput de rede ou ambos.

Serviços sensíveis à latência são fornecidos em locais de borda usando uma rede global da AWS, como o Amazon CloudFront. Esses locais de borda costumam oferecer serviços, como rede de entrega de conteúdo (CDN) e sistema de nomes de domínio (DNS). Ao ter esses serviços na borda, as workloads podem responder com baixa latência a solicitações de conteúdo ou resolução de DNS. Esses serviços também fornecem serviços geográficos, como direcionamento geográfico de conteúdo (fornecendo conteúdo diferente conforme o local do usuário final) ou encaminhamento por latência para direcionar os usuários finais à região mais próxima (latência mínima).

Use serviços de borda para reduzir a latência e possibilitar o armazenamento do conteúdo em cache. Configure corretamente o controle de cache para DNS e HTTP/HTTPS a fim de aproveitar ao máximo essas abordagens.

### Etapas da implementação

- Capture informações sobre o tráfego IP que entra e sai das interfaces de rede.
  - Registro em log do tráfego IP usando logs de fluxo de VPC
  - Como o endereço IP do cliente é preservado no AWS Global Accelerator

- Analise os padrões de acesso à rede em sua workload para identificar como os usuários utilizam sua aplicação.
  - Use ferramentas de monitoramento, como Amazon CloudWatch e o AWS CloudTrail, para coletar dados sobre as atividades da rede.
  - Analise os dados para identificar o padrão de acesso à rede.
- Selecione as Regiões para implantação da workload com base nos seguintes elementos fundamentais:
  - A localização dos seus dados: para aplicações com uso intenso de dados (como big data e machine learning), o código da aplicação deve ser executado o mais perto possível dos dados.
  - A localização dos seus usuários: para aplicações voltadas ao usuário, escolha uma Região (ou Regiões) próxima dos clientes de sua workload.
  - Outras restrições: leve em conta restrições, como custo e conformidade, conforme explicado em O que considerar ao selecionar uma região para suas workloads.
- Use zonas locais da AWS para executar workloads como renderização de vídeo. As zonas locais permitem que você se beneficie de ter recursos de computação e armazenamento mais próximos dos usuários finais.
- Use AWS Outposts para workloads que precisam permanecer on-premises e onde você deseja que essa workload seja executada ininterruptamente com o restante de suas workloads na AWS.
- Aplicações, como streaming de vídeo ao vivo em alta resolução, áudio de alta fidelidade ou realidade aumentada/realidade virtual (RA/RV), exigem latência ultrabaixa para dispositivos 5G. Para tais aplicações, considere o AWS WavelengthO AWS Wavelength incorpora serviços de armazenamento e computação da AWS em redes 5G, fornecendo a infraestrutura móvel de computação de borda para desenvolver, implantar e escalar aplicações de latência ultrabaixa.
- Use armazenamento em cache local ou soluções de armazenamento em cache da AWS para ativos usados com frequência a fim de aumentar a performance, reduzir a movimentação de dados e reduzir o impacto ambiental.

| Service | Quando usar |
| --- | --- |
| Amazon CloudFront | Use para armazenar conteúdo estático em cache, como imagens, scripts e vídeos, bem como conteúdo dinâmico, como respostas de API ou aplicações Web. |

| Service | Quando usar |
| --- | --- |
| Amazon ElastiCache | Use para armazenar conteúdo em cache para aplicações Web. |
| DynamoDB Accelerator | Use para adicionar aceleração na memória às suas tabelas do DynamoDB. |

- Use serviços que podem ajudar você a executar código mais perto dos usuários da workload, como a seguir:

| Serviço | Quando usar |
| --- | --- |
| Lambda@edge | Use para operações com uso intenso de computação que são iniciadas quando objetos não estão no cache. |
| Funções do Amazon CloudFront | Use para casos de uso simples, como solicitações HTTP(s) ou manipulações de resposta, que podem ser iniciadas por funções de curta duração. |
| AWS IoT Greengrass | Use para executar computação local, mensagens e armazenamento de dados em cache para dispositivos conectados. |

- Algumas aplicações exigem pontos de entrada fixos ou maior desempenho ao reduzir a tremulação e a latência de primeiro byte, além de aumentar o throughput. Essas aplicações podem se beneficiar de serviços de rede que fornecem endereços IP anycast estáticos e terminação TCP em locais da borda. AWS Global Accelerator pode melhorar o desempenho de suas aplicações em até 60% e fornecer failover rápido para arquiteturas multirregionais. O AWS Global Accelerator fornece endereços IP anycast estáticos que servem como um ponto de entrada fixo para suas aplicações hospedadas em uma ou mais Regiões da AWS. Esses endereços IP permitem que o tráfego entre na rede global da AWS o mais próximo possível dos usuários. O AWS Global Accelerator reduz o tempo de configuração da conexão inicial ao estabelecer uma conexão TCP entre o cliente e o local da borda da AWS mais próximo ao cliente. Analise o uso do AWS Global Accelerator para melhorar o desempenho das workloads de TCP/UDP e forneça failover rápido para arquiteturas de várias Regiões.

# Recursos

Práticas recomendadas relacionadas:

- COST07-BP02 Implementar regiões com base nos custos
- COST08-BP03 Implementar serviços para reduzir custos de transferência de dados
- REL10-BP01 Implantar a workload em vários locais
- REL10-BP02 Escolher os locais apropriados para sua implantação de vários locais
- SUS01-BP01 Escolher a região com base nos requisitos empresariais e nas metas de sustentabilidade
- SUS02-BP04 Otimizar o posicionamento geográfico das workloads com base nos respectivos requisitos de rede
- SUS04-BP07 Minimizar a movimentação de dados entre redes

Documentos relacionados:

- AWS Global Infrastructure (Infraestrutura global da AWS)
- AWS Local Zones and AWS Outposts, choosing the right technology for your edge workload (Zonas locais da AWS e AWS Outposts: como escolher a tecnologia certa para sua workload de borda)
- Grupos de posicionamento
- zonas locais da AWS
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Vídeos relacionados:

- AWS Local Zones Explainer Video (Vídeo de explicação de zonas locais da AWS)
- AWS Outposts: Overview and How it Works (AWS Outposts: visão geral e como funciona

- AWS re:Invent 2021 - AWS Outposts: Bringing the AWS experience on premises (AWS re:Invent 2021 - AWS Outposts: como trazer a experiência da AWS para ambientes on-premises)
- AWS re:Invent 2020: AWS Wavelength: Run apps with ultra-low latency at 5G edge (AWS re:Invent 2020: AWS Wavelength: execute aplicativos com latência ultrabaixa na borda 5G)
- AWS re:Invent 2022 - AWS Local Zones: Building applications for a distributed edge (AWS re:Invent 2022: zonas locais da AWS: como criar aplicações para uma borda distribuída)
- AWS re:Invent 2021 - Building low-latency websites with Amazon CloudFront (AWS re:Invent 2021: criação de sites de baixa latência com o Amazon CloudFront)
- AWS re:Invent 2022 - Improve performance and availability with AWS Global Accelerator (AWS re:Invent 2022: melhore a performance e a disponibilidade com o AWS Global Accelerator)
- AWS re:Invent 2022 - Build your global wide area network using AWS (AWS re:Invent 2022: crie sua rede de longa distância usando a AWS)
- AWS re:Invent 2020: Global traffic management with Amazon Route 53 (AWS re:Invent 2020: gerenciamento de tráfego global com o Amazon Route 53)

Exemplos relacionados:

- Workshop do AWS Global Accelerator
- Handling Rewrites and Redirects using Edge Functions (Lidar com reescritas e redirecionamentos usando funções da borda)

## PERF04-BP07 Otimizar a configuração da rede com base em métricas

Use dados coletados e analisados para tomar decisões bem informadas sobre a otimização da configuração da rede.

Antipadrões comuns:

- Você pressupõe que todos os problemas relacionados à performance são relacionados à aplicação.
- Você só testa a performance da rede a partir de um local próximo ao local em que implantou a carga de trabalho.
- Você usa configurações-padrão para todos os serviços de rede.

- Você provisiona em excesso recursos de rede para fornecer capacidade suficiente.

Benefícios de estabelecer esta prática recomendada: coletar as métricas necessárias da rede da AWS e implementar ferramentas de monitoramento de rede permite entender o desempenho da rede e otimizar as configurações dela.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

## Orientação para implementação

Monitorar o tráfego de entrada e saída das VPCs, sub-redes ou interfaces de rede é fundamental para entender como utilizar os recursos de rede da AWS e otimizar as configurações da rede. Ao usar as ferramentas de rede da AWS a seguir, é possível verificar mais informações sobre o uso do tráfego, o acesso à rede e os logs.

### Etapas da implementação

- Identifique as principais métricas de desempenho, como latência ou perda de pacotes. A AWS fornece diversas ferramentas que podem ajudar você a coletar essas métricas. Ao usar as ferramentas a seguir, é possível verificar mais informações sobre o uso do tráfego, o acesso à rede e os logs:

| Ferramenta da AWS | Onde usar |
| --- | --- |
| Amazon VPC IP Address Manager. | Use o IPAM para planejar, rastrear e monitorar endereços IP para workloads da AWS e on-premises. Essa é uma prática recomendada para otimizar o uso e a alocação de endereços IP. |
| Logs de fluxo da VPC | Use os logs de fluxo da VPC para obter informações detalhadas sobre o tráfego de entrada e saída das interfaces de rede nas VPCs. Com os logs de fluxo da VPC, é possível diagnosticar regras extremame nte restritivas ou permissivas do grupo de segurança e determinar a direção do tráfego de entrada e saída das interfaces de rede. |

| Ferramenta da AWS | Onde usar |
| --- | --- |
| Logs de fluxo do AWS Transit Gateway | Use logs de fluxo do AWS Transit Gateway para capturar informações sobre o tráfego IP que entra e sai dos seus gateways de trânsito. |
| Registro em log de consultas ao DNS | Registre informações sobre consultas ao DNS, públicas ou privadas, que o Route 53 recebe. Com os logs de DNS, é possível otimizar as configurações de DNS entendend o o domínio ou subdomínio solicitado ou os locais da borda do Route 53 que responderam às consultas ao DNS. |
| Reachability Analyzer | O Reachability Analyzer ajuda você a analisar e depurar a capacidade de alcance da rede. O Reachability Analyzer é uma ferramenta de análise de configuração que permite realizar testes de conectividade entre um recurso da origem e um do destino nas VPCs. Essa ferramenta ajuda a verificar se a configura ção da rede corresponde à conectividade pretendida. |
| Network Access Analyzer | O Network Access Analyzer ajuda você a entender o acesso da rede aos seus recursos. É possível usar o Network Access Analyzer para especificar os requisitos de acesso à rede e identificar possíveis caminhos de rede que não atendem aos requisitos especific ados. Ao otimizar a configuração da rede correspondente, é possível entender e verificar o estado da rede e demonstrar se a rede na AWS atende aos seus requisitos de conformidade. |

| Ferramenta da AWS | Onde usar |
| --- | --- |
| Amazon CloudWatch | Use o Amazon CloudWatch e ative as métricas apropriadas para as opções de rede. Escolha a métrica de rede certa para sua workload. Por exemplo, é possível habilitar métricas para o uso do endereço de rede da VPC, o gateway NAT da VPC, o AWS Transit Gateway, o túnel da VPN, o AWS Network Firewall, o Elastic Load Balancing e o AWS Direct Connect. Monitorar continuamente as métricas é uma prática recomendada para observar e entender o status e o uso da rede, o que ajuda a otimizar a configuração da rede com base em suas observações. |
| AWS Network Manager | Usando o AWS Network Manager, você pode monitorar o desempenho histórico e em tempo real da rede global da AWS para fins operacionais e de planejamento. O Network Manager fornece latência de rede agregada entre as Regiões da AWS e Zonas de Disponibilidade e dentro de cada Zona de Disponibilidade, permitindo que você entenda melhor como o desempenho da sua aplicação se relaciona com o desempenho da rede da AWS subjacente. |
| Amazon CloudWatch RUM | Use o Amazon CloudWatch RUM para coletar as métricas que fornecem os insights que ajudam a identificar, entender e melhorar a experiência do usuário. |

- Identifique os principais interlocutores e os padrões de tráfego de aplicações usando VPC e logs de fluxo do AWS Transit Gateway.

- Avalie e otimize sua arquitetura de rede atual, incluindo VPCs, sub-redes e roteamento. Como exemplo, você pode avaliar como diferentes emparelhamentos de VPC ou AWS Transit Gateway podem ajudar a melhorar a rede em sua arquitetura.
- Avalie os caminhos de roteamento em sua rede para verificar se o caminho mais curto entre os destinos é sempre usado. O Network Access Analyzer pode ajudar nessa tarefa.

## Recursos

Documentos relacionados:

- Logs de fluxo da VPC
- Registro em log de consulta ao DNS público
- O que é o IPAM?
- O que é o Reachability Analyzer?
- O que é o Network Access Analyzer?
- Métricas do CloudWatch para suas VPCs
- Otimize o desempenho e reduza os custos de análise da rede com os logs de fluxo da VPC no formato Apache Parquet
- Monitoramento de suas redes globais e principais com métricas do Amazon CloudWatch
- Monitore continuamente o tráfego e os recursos da rede

Vídeos relacionados:

- Networking best practices and tips with the AWS Well-Architected Framework (Práticas recomendadas e dicas de redes com o AWS Well-Architected Framework)
- Monitoring and troubleshooting network traffic (Monitoramento e resolução de problemas de tráfego de rede)

Exemplos relacionados:

- Workshops de redes da AWS
- Monitoramento de rede da AWS

# Processo e cultura

Ao arquitetar workloads, há princípios e práticas que você pode adotar para ajudar na melhor execução de workloads de nuvem eficientes e de alto desempenho. Essa área de foco oferece as práticas recomendadas para ajudar a adotar uma cultura que promova a eficiência do desempenho das workloads na nuvem.

Considere estes princípios fundamentais para construir essa cultura:

- Infraestrutura como código: defina sua infraestrutura como código usando abordagens como modelos do AWS CloudFormation. O uso de modelos permite que você coloque a infraestrutura no controle de origem junto com o código e as configurações de sua aplicação. Isso permite aplicar à sua infraestrutura as mesmas práticas usadas para desenvolver software, possibilitando uma iteração rápida.
- Pipeline de implantação: use um pipeline de integração/implantação contínuas (CI/CD) (por exemplo, repositório de código-fonte, sistemas de compilação, implantação e automação de teste) para implantar sua infraestrutura. Isso permite que você implante de maneira repetível, consistente e econômica enquanto itera.
- Métricas bem definidas: configure e monitore suas métricas para capturar os indicadores-chave de performance (KPIs). Recomendamos o uso tanto de métricas técnicas quanto de negócios. Para aplicativos móveis ou sites, as principais métricas são a captura do tempo até o primeiro byte ou renderização. Outras métricas geralmente aplicáveis incluem contagem de thread, taxa de coleta de resíduos e estados de espera. Métricas de negócio, como o custo cumulativo agregado por solicitação, podem alertá-lo sobre maneiras de reduzir os custos. Considere com cuidado como você planeja interpretar as métricas. Por exemplo, você poderia escolher o máximo ou o 99.º percentil, em vez da média.
- Testar a performance automaticamente: como parte de seu processo de implantação, comece automaticamente testes de performance após a aprovação bem-sucedida dos testes de execução mais rápidos. A automação deve criar um novo ambiente, configurar as condições iniciais, como dados de teste, e então executar uma série de benchmarks e testes de carga. Os resultados desses testes então devem ser vinculados de volta à compilação para que você possa acompanhar as mudanças de performance ao longo do tempo. Para testes de execução longa, você pode tornar essa parte do pipeline assíncrona do restante da compilação. Como alternativa, você pode realizar testes de performance durante a noite usando instâncias spot do Amazon EC2.
- Geração de carga: você deve criar uma série de scripts de teste que repliquem jornadas sintéticas ou pré-gravadas do usuário. Esses scripts devem ser idempotentes e não acoplados, e talvez seja

necessário incluir scripts de pré-aquecimento para gerar resultados válidos. Seus scripts de teste devem replicar tanto quanto for possível o comportamento do uso na produção. É possível usar soluções de software ou Software como Serviço (SaaS) para gerar a carga. Considere o uso de soluções do AWS Marketplace e instâncias spot. Elas podem ser maneiras econômicas de gerar a carga.

- Visibilidade da performance: as métricas principais devem estar visíveis à sua equipe, especialmente métricas relacionadas a cada versão de build. Isso permite que você veja qualquer tendência positiva ou negativa importante ao longo do tempo. Você também deve exibir métricas do número de erros ou exceções para garantir que esteja testando um sistema em funcionamento.
- Visualização: use técnicas de visualização que deixem claro onde os problemas de performance, hot spots, estados de espera ou baixa utilização estão ocorrendo. Sobreponha métricas de performance a diagramas de arquitetura – código ou gráficos de chamada podem ajudar a identificar problemas rapidamente.
- Processo de revisão regular: arquiteturas com baixa performance geralmente são o resultado de um processo de análise de performance inexistente ou problemático. Se sua arquitetura está funcionando mal, a implementação de um processo de análise de desempenho permite que você promova melhorias iterativas.
- Otimização contínua: adote uma cultura para otimizar continuamente a eficiência do desempenho da workload na nuvem.

Práticas recomendadas

- PERF05-BP01 Estabeleça indicadores-chave de desempenho (KPIs) para medir a integridade e o desempenho da workload
- PERF05-BP02 Use soluções de monitoramento para entender as áreas em que o desempenho é mais crítico
- PERF05-BP03 Defina um processo para melhorar a performance da workload
- PERF05-BP04 Faça o teste de carga da workload
- PERF05-BP05 Use a automação para corrigir proativamente problemas relacionados ao desempenho
- PERF05-BP06 Mantenha a workload e os serviços atualizados
- PERF05-BP07 Analise as métricas regularmente

# PERF05-BP01 Estabeleça indicadores-chave de desempenho (KPIs) para medir a integridade e o desempenho da workload

Identifique os KPIs que medem o desempenho da workload de forma quantitativa e qualitativa. Os KPIs ajudam você a medir a integridade e o desempenho de uma workload relacionada a uma meta empresarial.

Antipadrões comuns:

- Você só monitora as métricas no nível do sistema para obter informações da workload e não compreende aos impactos dessas métricas nos negócios.
- Você pressupõe que os KPIs já estejam publicados e compartilhados como dados de métricas comuns.
- Você não define um KPI quantitativo e mensurável.
- Você não alinha os KPIs às metas ou estratégias empresariais.

Benefícios de estabelecer esta prática recomendada: Identificar KPIs específicos que representam a integridade e o desempenho da workload ajuda a alinhar as equipes em suas prioridades e a definir resultados empresariais bem-sucedidos. O compartilhamento dessas métricas com todos os departamentos fornece visibilidade e alinhamento dos limites, das expectativas e do impacto nos negócios.

Nível de risco exposto se esta prática recomendada não for estabelecida: alto

## Orientação para implementação

Os KPIs permitem que as empresas e as equipes de engenharia alinhem a medição das metas e estratégias de como esses fatores são combinados para produzir resultados empresariais. Por exemplo, a workload de um site pode usar o tempo de carregamento da página como uma indicação de desempenho geral. Essa métrica seria um dos vários pontos de dados que medem a experiência do usuário. Além de identificar os limites do tempo de carregamento da página, documente o resultado esperado ou o risco da empresa se o desempenho ideal não for atingido. Um longo tempo de carregamento da página afeta diretamente os usuários finais, diminui a classificação da experiência do usuário e pode resultar em perda de clientes. Ao definir os limites dos KPIs, combine os testes comparativos do setor e as expectativas dos usuários finais. Por exemplo, se o teste comparativo do setor atual for o carregamento de uma página da web em dois segundos, mas os

usuários finais esperarem que uma página da web seja carregada em um segundo, você deverá pensar nos dois pontos de dados ao estabelecer o KPI.

Sua equipe deve avaliar os KPIs da workload usando dados detalhados em tempo real e dados históricos para referência, e criar painéis que calculem as métricas nos dados de KPI para derivar informações operacionais e de utilização. Os KPIs devem ser documentados e incluir limites que apoiem as metas e estratégias empresariais, bem como mapeados de acordo com as métricas que estão sendo monitoradas. Os KPIs devem ser revisitados quando mudam as metas e as estratégias da empresa ou os requisitos dos usuários finais.

## Etapas da implementação

1. Identifique e documente as principais partes interessadas da empresa.
2. Trabalhe com essas partes interessadas para definir e documentar os objetivos da workload.
3. Analise as práticas recomendadas do setor para identificar KPIs relevantes alinhados aos objetivos da workload.
4. Use as práticas recomendadas do setor e os objetivos da workload para definir metas de KPI da workload. Use essas informações para definir limites de KPI no nível de gravidade ou de alarme.
5. Identifique e documente o risco e o impacto no caso de um KPI não ser atendido.
6. Identifique e documente métricas que podem ajudar a estabelecer os KPIs.
7. Use ferramentas de monitoramento, como Amazon CloudWatch ou AWS Config para coletar métricas e medir KPIs.
8. Use painéis para visualizar e comunicar os KPIs com as partes interessadas.
9. Revise e analise regularmente as métricas para identificar áreas da workload que precisam ser aprimoradas.
10. Revise os KPIs quando as metas empresariais ou a performance da workload mudarem.

## Recursos

Documentos relacionados:

- Documentação da CloudWatch
- AWS Partners de monitoramento, registro em log e performance
- Documentação do X-Ray
- Using Amazon CloudWatch dashboards

- Amazon QuickSight KPIs

Vídeos relacionados:

- AWS re:Invent 2019: Scaling up to your first 10 million users
- Cut through the chaos: Gain operational visibility and insight
- Build a monitoring plan

Exemplos relacionados:

- Creating a dashboard with Amazon QuickSight

## PERF05-BP02 Use soluções de monitoramento para entender as áreas em que o desempenho é mais crítico

Entenda e identifique áreas em que aumentar a performance de sua workload causará um impacto positivo sobre a eficiência ou a experiência do cliente. Por exemplo, um site que tenha muita interação com o cliente se beneficiaria do uso de serviços de borda para aproximar a entrega de conteúdo dos clientes.

Antipadrões comuns:

- Você pressupõe que as métricas de computação padrão, como utilização de CPU ou pressão de memória, são suficientes para detectar problemas de performance.
- Você só usa as métricas comuns registradas pelo software de monitoramento selecionado.
- Você só revisa as métricas quando há um problema.

Benefícios de estabelecer esta prática recomendada: Compreender áreas críticas de desempenho ajuda os proprietários de workloads a monitorar KPIs e priorizar melhorias de alto impacto.

Nível de risco exposto se essa prática recomendada não for estabelecida: alto

### Orientação para implementação

Configure um rastreamento completo para identificar padrões de tráfego, latência e áreas de desempenho críticas. Monitore os padrões de acesso aos dados para consultas lentas ou dados

particionados e fragmentados incorretamente. Identifique as áreas de restrição da workload usando o teste ou monitoramento de carga.

aumentar a eficiência do desempenho entendendo sua arquitetura, os padrões de tráfego e os padrões de acesso aos dados, além de identificar os tempos de latência e processamento. Identificar possíveis gargalos que possam afetar a experiência do cliente com o crescimento da workload. Depois de investigar essas áreas, veja qual solução você pode implantar para eliminar esses problemas de desempenho.

## Etapas da implementação

1. Configure um monitoramento completo para capturar todos os componentes e as métricas da workload. Aqui estão alguns exemplos de soluções de monitoramento na AWS.

| Service | Onde usar |
| --- | --- |
| Monitoramento de usuários reais (RUM) do Amazon CloudWatch | para capturar as métricas de performance da aplicação de sessões de front-end e do lado do cliente de usuários reais. |
| AWS X-Ray | para monitorar o tráfego por meio das camadas de aplicação e identificar a latência entre componentes e dependências. Use os mapas do serviço X-Ray para ver os relaciona mentos e a latência entre os componentes da workload. |
| Insights de Performance do Amazon Relational Database Service | Para ver as métricas de performance do banco de dados e identificar melhorias de performance. |
| Monitoramento avançado do Amazon RDS | Para ver métricas de performance do SO do banco de dados. |
| Amazon DevOps Guru | Para detectar padrões operacionais anormais a fim de que você possa identificar problemas operacionais antes que eles afetem os clientes. |

2. Realize testes para gerar métricas, identificar padrões de tráfego, gargalos e áreas de desempenho críticas. Aqui estão alguns exemplos de como realizar testes:
   - Configure o Canários sintéticos do CloudWatch para imitar programaticamente as atividades do usuário baseadas no navegador usando trabalhos cron do Linux ou expressões de taxa para gerar métricas consistentes ao longo do tempo.
   - Use o Testes de carga distribuída da AWS para gerar tráfego de pico ou testar a workload na taxa de crescimento esperada.
3. Avalie as métricas e a telemetria para identificar as áreas de desempenho críticas. Avalie essas áreas com sua equipe para discutir sobre o monitoramento e as soluções visando evitar gargalos.
4. Experimente melhorias de desempenho e meça essas alterações com dados. Por exemplo, você pode usar o CloudWatch Evidently para testar novas melhorias e impactos na performance da workload.

## Recursos

Documentos relacionados:

- Amazon Builders' Library
- Documentação do X-Ray
- Amazon CloudWatch RUM
- Amazon DevOps Guru

Vídeos relacionados:

- The Amazon Builders' Library: 25 years of Amazon operational excellence
- Visual Monitoring of Applications with Amazon CloudWatch Synthetics

Exemplos relacionados:

- Measure page load time with Amazon CloudWatch Synthetics (Medição do tempo de carga da página com o Amazon CloudWatch Synthetics)
- Amazon CloudWatch RUM Web Client (Cliente da web do Amazon CloudWatch RUM)
- X-Ray SDK para Node.js
- X-Ray SDK para Python

- X-Ray SDK para Java
- X-Ray SDK para .Net
- X-Ray SDK para Ruby
- Daemon do X-Ray
- Testes de carga distribuída na AWS

# PERF05-BP03 Defina um processo para melhorar a performance da workload

Defina um processo para avaliar novos serviços, padrões de design, tipos de recursos e configurações conforme ficarem disponíveis. Por exemplo, execute testes de performance existentes em novas ofertas de instância para determinar o potencial delas de aprimorar sua carga de trabalho.

Antipadrões comuns:

- Você pressupõe que sua arquitetura atual é estática e não será atualizada ao longo do tempo.
- Você apresenta alterações de arquitetura ao longo do tempo sem justificativa de métrica.

Benefícios de estabelecer esta prática recomendada: Ao definir seu processo para fazer alterações de arquitetura, é possível usar os dados coletados para influenciar o projeto da workload ao longo do tempo.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

A performance de sua carga de trabalho tem algumas restrições importantes. Guarde essas restrições para saber que tipos de inovação podem aumentar a performance de sua carga de trabalho. Use essas informações enquanto estiver aprendendo sobre novos serviços ou tecnologias que surgem e identificar maneiras de reduzir restrições ou gargalos.

Identifique as principais restrições de desempenho da workload. Documente suas restrições de performance da carga de trabalho para que você saiba quais tipos de inovação podem aprimorar a performance da carga de trabalho.

### Etapas da implementação

- Identifique seus KPIs de performance da workload conforme descrito em PERF05-BP01 Estabeleça indicadores-chave de desempenho (KPIs) para medir a integridade e o desempenho da workload para basear sua workload.
- Use Ferramentas de observabilidade da AWS para coletar métricas de performance e medir KPIs.
- Faça uma análise aprofundada para identificar as áreas (como configuração e código da aplicação) na workload que estão com baixa performance, conforme descrito em PERF05-BP02 Use soluções de monitoramento para entender as áreas em que o desempenho é mais crítico.
- Use suas ferramentas de análise e desempenho para identificar a estratégia de otimização de desempenho.
- Use ambientes de sandbox ou de pré-produção para validar a eficácia da estratégia.
- Implemente as mudanças na produção e monitore constantemente o desempenho da workload.
- Documente as melhorias e comunique isso às partes interessadas.

## Recursos

Documentos relacionados:

- Blog da AWS
- Novidades da AWS

Vídeos relacionados:

- Canal AWS Events no YouTube
- Canal Online Tech Talks da AWS no YouTube
- Canal da Amazon Web Services no YouTube

Exemplos relacionados:

- AWS Github
- AWS Skill Builder

# PERF05-BP04 Faça o teste de carga da workload

Teste sua workload para verificar se ela pode lidar com a carga de produção e identificar qualquer gargalo de desempenho.

Antipadrões comuns:

- Você realiza um teste de carga de peças individuais da workload, mas não toda a workload.
- Você realiza um teste de carga em uma infraestrutura que não é igual ao seu ambiente de produção.
- Você só realiza testes de carga para a carga esperada e não para além dela, para ajudar a prever onde você pode ter problemas futuros.
- Você realiza testes de carga sem consultar a política de testes do Amazon EC2 e enviar um formulário de envio de eventos simulados. Isso faz com que o teste não seja executado, pois parece um evento de negação de serviço.

Benefícios de estabelecer esta prática recomendada: Medir sua performance em um teste de carga mostrará onde você será afetado à medida que a carga aumentar. Com isso você terá a capacidade de antecipar as alterações necessárias antes que elas afetem sua carga de trabalho.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

## Orientação para implementação

O teste de carga na nuvem é um processo para medir o desempenho da workload na nuvem em condições realistas com a carga esperada do usuário. Esse processo envolve o provisionamento de um ambiente de nuvem semelhante ao de produção, o uso de ferramentas de teste de carga para gerar carga e a análise de métricas para avaliar a capacidade da workload de lidar com cargas realistas. Execute os testes de carga usando versões sintéticas ou limpas dos dados de produção (remova informações confidenciais ou de identificação). Realize testes de carga automaticamente como parte de seu pipeline de entrega e compare os resultados a Key Performance Indicators (KPI – Indicadores-chave de performance) e limites predefinidos. Esse processo ajuda você a continuar alcançando o desempenho necessário.

### Etapas da implementação

- Configure o ambiente de teste com base no ambiente de produção. É possível usar os serviços da AWS para executar ambientes em escala de produção para testar a arquitetura.

- Escolha e configure a ferramenta de teste de carga adequada à workload.
- Defina os cenários e parâmetros do teste de carga (como duração do teste e número de usuários).
- Execute cenários de teste em grande escala. Aproveite a Nuvem AWS para testar a workload e descobrir se há uma falha na escala ou se ela está com a escala reduzida horizontalmente de maneira não linear. Por exemplo, use instâncias spot para gerar cargas a um baixo custo e descobrir gargalos antes que eles ocorram em produção.
- Monitore e registre métricas de desempenho (como throughput e tempo de resposta). O Amazon CloudWatch pode coletar métricas entre os recursos em sua arquitetura. Você também pode coletar e publicar métricas personalizadas para descobrir métricas de negócio ou derivadas.
- Analise os resultados para identificar gargalos de desempenho e áreas para melhorias.
- Documente e relate o processo e os resultados do teste de carga.

## Recursos

Documentos relacionados:

- AWS CloudFormation
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Testes de carga distribuída na AWS

Vídeos relacionados:

- Solving with AWS Solutions: Distributed Load Testing

- Optimize applications through Amazon CloudWatch RUM
- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Testes de carga distribuída na AWS

# PERF05-BP05 Use a automação para corrigir proativamente problemas relacionados ao desempenho

Use os indicadores-chave de performance (KPIs), aliados a sistemas de monitoramento e alerta, para abordar proativamente problemas relacionados à performance.

Antipadrões comuns:

- Você só permite que a equipe de operações faça alterações operacionais na workload.
- Você permite todos os filtros de alarmes para a equipe de operações, sem correção proativa.

Benefícios de estabelecer esta prática recomendada: A correção proativa de ações de alarme permite que a equipe de suporte se concentre nos itens que não são acionáveis automaticamente. Isso ajuda a equipe de operações a lidar com todos os alarmes sem ficar sobrecarregada e, em vez disso, se concentrar apenas nos alarmes críticos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

## Orientação para implementação

Sempre que possível, use alarmes para desencadear ações automatizadas visando corrigir problemas. Se a resposta automatizada não for possível, encaminhe o alarme para aqueles capazes de responder. Por exemplo, você pode ter um sistema capaz de prever os valores de indicadores-chave de desempenho (KPI) esperados e emitir um alarme quando eles ultrapassarem determinados limites, ou uma ferramenta capaz de interromper ou reverter automaticamente as implantações caso os KPIs estejam fora dos valores esperados.

Implemente processos que deem visibilidade à performance conforme sua carga de trabalho estiver sendo executada. Para determinar se a performance da carga de trabalho é ideal, crie painéis de monitoramento e estabeleça normas de linha de base para as expectativas de performance.

### Etapas da implementação

- Identifique e compreenda o problema de desempenho que pode ser corrigido automaticamente. Use soluções de monitoramento da AWS, como o Amazon CloudWatch ou o AWS X-Ray, para ajudar você a entender melhor a causa raiz do problema.
- Crie um plano e um processo de correção detalhados que possam ser usados para corrigir automaticamente o problema.

- Configure o gatilho para iniciar automaticamente o processo de correção. Por exemplo, você pode definir um acionador para reiniciar automaticamente uma instância quando ela atinge determinado limite de utilização da CPU.
- Use serviços e tecnologias da AWS para automatizar o processo de correção. Por exemplo: AWS Systems Manager Automation fornece uma maneira segura e escalável de automatizar o processo de correção.
- Teste o processo de correção automatizado em um ambiente de pré-produção.
- Após o teste, implemente o processo de correção no ambiente de produção e monitore constantemente para identificar áreas de melhoria.

## Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Parceiros da AWS Partner Network de monitoramento, registro em log e performance
- Documentação do X-Ray
- Using Alarms and Alarm Actions in CloudWatch

Vídeos relacionados:

- Intelligently automating cloud operations (Automatizar de forma inteligente as operações na nuvem)
- Setting up controls at scale in your AWS environment
- Automating patch management and compliance using AWS
- How Amazon uses better metrics for improved website performance (Como a Amazon usa métricas melhores para aprimorar o desempenho do site)

Exemplos relacionados:

- CloudWatch Logs Customize Alarms

# PERF05-BP06 Mantenha a workload e os serviços atualizados

Atualize-se com relação aos novos serviços e atributos de nuvem para adotar recursos eficientes, remover problemas e melhorar a eficiência geral do desempenho da workload.

Antipadrões comuns:

- Você pressupõe que sua arquitetura atual é estática e não será atualizada ao longo do tempo.
- Você não tem nenhum sistema ou ritmo regular para avaliar se software ou pacotes atualizados são compatíveis com sua workload.

Benefícios de estabelecer esta prática recomendada: Ao estabelecer um processo para se atualizar sobre novos serviços e ofertas, você pode adotar novos atributos e recursos, resolver problemas e melhorar a performance da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

## Orientação para implementação

Avalie maneiras de melhorar o desempenho conforme são disponibilizados novos serviços, padrões de design e atributos de produtos. Determine quais deles poderiam aprimorar o desempenho ou aumentar a eficiência da workload por meio de avaliações, discussões internas ou análises externas. Defina um processo para avaliar atualizações, novos recursos e serviços relevantes para sua workload. Por exemplo, crie uma prova de conceito que use novas tecnologias ou consulte um grupo interno. Ao testar novas ideias ou serviços, execute testes de desempenho para medir o impacto que eles têm sobre o desempenho da workload.

## Etapas da implementação

- Fazer o inventário de software e arquitetura da workload e identificar os componentes que precisam ser atualizados.
- Identifique novidades e atualize fontes relacionadas aos componentes da workload. Por exemplo, você pode se inscrever no What's New at AWS para os produtos que correspondem ao componente da workload. Você pode assinar o feed RSS ou gerenciar suas assinaturas de e-mail.
- Defina um cronograma para avaliar novos serviços e atributos para a workload.
  - Você pode usar o inventário do AWS Systems Manager para coletar metadados de sistema operacional (SO), aplicação e instância das instâncias do Amazon EC2 e entender rapidamente

quais instâncias executam o software e as configurações exigidas pela política de software e quais instâncias precisam ser atualizadas.

- Entenda como atualizar os componentes de sua workload. Aproveite a agilidade da nuvem para testar rapidamente como novos atributos podem melhorar a workload com o intuito de obter eficiências de performance.
- Use automação no processo de atualização para reduzir o nível de esforço para implantar novos recursos e limitar erros causados por processos manuais.
  - Você pode usar o CI/CD para atualizar automaticamente AMIs, imagens de contêiner e outros artefatos relacionados à aplicação de nuvem.
  - Você pode usar ferramentas, como AWS Systems Manager Patch Manager para automatizar o processo de atualizações do sistema e programar a atividade usando Janelas de Manutenção do AWS Systems Manager.
- Documente seu processo para avaliar atualizações e novos serviços. Forneça aos proprietários o tempo e o espaço necessários para pesquisar, testar, experimentar e validar atualizações e novos serviços. Consulte novamente os KPIs e requisitos empresariais documentados para ajudar a priorizar qual atualização trará um impacto positivo à empresa.

## Recursos

Documentos relacionados:

- Blog da AWS
- Novidades da AWS

Vídeos relacionados:

- Canal AWS Events no YouTube
- Canal Online Tech Talks da AWS no YouTube
- Canal da Amazon Web Services no YouTube

Exemplos relacionados:

- Laboratórios do Well-Architected: Gerenciamento de inventário e patches
- Laboratório: AWS Systems Manager

# PERF05-BP07 Analise as métricas regularmente

Como parte da manutenção de rotina, ou em resposta a eventos ou incidentes, analise as métricas que são coletadas. Use essas análises para identificar quais métricas foram essenciais para resolver problemas e quais métricas adicionais poderiam ajudar a identificar, resolver ou prevenir problemas se estivessem sendo acompanhadas.

Antipadrões comuns:

- Você permite que as métricas permaneçam em um estado de alarme por um período prolongado.
- Você cria alarmes que não são acionáveis por um sistema de automação.

Benefícios de estabelecer esta prática recomendada: Analise continuamente as métricas que estão sendo coletadas para garantir que identifiquem, resolvam ou evitem problemas corretamente. As métricas também podem se tornar obsoletas se você permitir que elas permaneçam em um estado de alarme por um período prolongado.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

## Orientação para implementação

Melhore constantemente a coleta e o monitoramento de métricas. Como parte da resposta a incidentes ou eventos, avalie as métricas que foram úteis para resolver o problema e quais poderiam ter ajudado, mas não estão sendo acompanhadas no momento. Use este método para aprimorar a qualidade das métricas coletadas, de modo que você possa prevenir ou resolver incidentes futuros mais rapidamente.

Como parte da resposta a incidentes ou eventos, avalie as métricas que foram úteis para resolver o problema e quais poderiam ter ajudado, mas não estão sendo acompanhadas no momento. Use esses dados para aprimorar a qualidade das métricas coletadas, de modo que você possa prevenir ou resolver incidentes futuros mais rapidamente.

### Etapas da implementação

1. Defina métricas essenciais de desempenho a serem monitoradas que estejam alinhadas ao seu objetivo de workload.
2. Defina uma linha de base e um valor desejável para cada métrica.
3. Defina uma frequência (como semanal ou mensal) para revisar as métricas essenciais.

4. Durante cada revisão, avalie as tendências e o desvio dos valores base. Procure gargalos ou anomalias de desempenho.
5. Para os problemas identificados, realize uma análise aprofundada da causa raiz para entender o principal motivo do problema.
6. Documente as descobertas e use estratégias para lidar com os problemas e gargalos identificados.
7. Avalie e melhore constantemente o processo de revisão de métricas.

## Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Collect metrics and logs from Amazon EC2 Instances and on-premises servers with the CloudWatch Agent
- Parceiros da AWS Partner Network de monitoramento, registro em log e performance
- Documentação do X-Ray

Vídeos relacionados:

- Setting up controls at scale in your AWS environment
- How Amazon uses better metrics for improved website performance (Como a Amazon usa métricas melhores para aprimorar o desempenho do site)

Exemplos relacionados:

- Creating a dashboard with Amazon QuickSight
- Level 100: Monitoring with CloudWatch Dashboards

# Conclusão

Atingir e manter a eficiência de performance requer uma abordagem conduzida por dados. Você deve avaliar ativamente os padrões de acesso e as concessões que viabilizarão a otimização para uma maior performance. O uso de um processo de análise baseado em benchmarks e testes de carga permite que você selecione os tipos de recursos e as configurações adequados. Tratar sua infraestrutura como código permite que você promova avanços em sua arquitetura de modo rápido e seguro, enquanto usa dados para tomar decisões baseadas em fatos sobre sua arquitetura. O estabelecimento de uma combinação de monitoramentos ativo e passivo garante que a performance de sua arquitetura não apresente degradação ao longo do tempo.

A AWS não mede esforços para ajudar você a criar arquiteturas que ofereçam uma performance eficiente enquanto entregam valor empresarial. Use as ferramentas e técnicas abordadas neste documento para garantir o sucesso.

# Colaboradores

Os indivíduos e empresas a seguir contribuíram para este documento:

- Sam Mokhtari, arquiteto sênior de soluções de eficiência da Amazon Web Services
- Josh Hart, arquiteto de soluções, Amazon Web Services
- Richard Trabing, arquiteto de soluções, Amazon Web Services
- Brett Looney, arquiteto-chefe de soluções da Amazon Web Services
- Nina Vogl, arquiteta-chefe de soluções da Amazon Web Services
- Eric Pullen, arquiteto de soluções, Amazon Web Services
- Julien Lépine, gerente especialista de SA, Amazon Web Services
- Ronnen Slasky, arquiteto de soluções, Amazon Web Services

# Leitura adicional

Para obter ajuda adicional, consulte as seguintes fontes:

- AWS Well-Architected Framework
- Centro de Arquitetura da AWS

# Revisões do documento

Para ser notificado sobre atualizações deste whitepaper, inscreva-se no RSS feed.

| Alteração | Descrição | Data |
| --- | --- | --- |
| Atualização e reestruturação importantes | O pilar foi reestruturado para ter cinco áreas de práticas recomendadas (antes eram oito). O conteúdo foi consolidado nas cinco áreas e atualizado.<br><br>As novas áreas de práticas recomendadas são Seleção de arquitetura, Computação e hardware, Gerenciamento de dados, Rede e entrega de conteúdoe Processo e cultura. | October 3, 2023 |
| Atualização secundária | Remoção de linguagem não inclusiva. | April 13, 2023 |
| Atualizações para o novo Framework | Práticas recomendadas atualizadas com orientações prescritivas e novas práticas recomendadas adicionadas. | April 10, 2023 |
| Whitepaper atualizado | Práticas recomendadas atualizadas com novas orientações para implementação. | December 15, 2022 |
| Whitepaper atualizado | Práticas recomendadas ampliadas e planos de melhoria adicionados. | October 20, 2022 |

| | | |
|---|---|---|
| Atualização secundária | Remoção de linguagem não inclusiva. | April 22, 2022 |
| Atualização secundária | Pilar Sustentabilidade adicionado à introdução. | December 2, 2021 |
| Atualizações secundárias | Links atualizados. | March 10, 2021 |
| Atualizações secundárias | Alteração do tempo limite do AWS Lambda para 900 segundos e correção do nome do Amazon Keyspaces (for Apache Cassandra). | October 5, 2020 |
| Atualização secundária | Link quebrado corrigido. | July 15, 2020 |
| Atualizações para a nova estrutura de trabalho | Análise e atualização importantes de conteúdo | July 8, 2020 |
| Whitepaper atualizado | Pequena atualização devido a problemas gramaticais | July 1, 2018 |
| Whitepaper atualizado | Whitepaper atualizado para refletir as alterações na AWS | November 1, 2017 |
| Publicação inicial | Publicação do pilar Eficiência de performance: AWS Well-Architected Framework. | November 1, 2016 |